GENESET MAPR: CHARACTERIZATION OF GENE SETS THROUGH HETEROGENEOUS NETWORK PATTERNS

BY

GREGORY LINKOWSKI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Associate Professor Shobha Vasudevan

ABSTRACT

Often, machine learning and big data concepts are applied to problems without a proper appreciation of their limitations or domain context. At the same time there is a growing appreciation for the ability of networks to represent more complex connections between data points than previous structures. However, established machine learning approaches rarely take advantage of such structures and must be adapted. We present here a method that utilizes patterns of connections within heterogeneous networks to score items by their similarity to an input set. We apply the idea of meta-paths as an abstraction to counteract typical big data problems of noise and overfitting. We also aim to demystify the black-box nature of machine learning by providing intuitive feedback about why items are considered similar. While the method presented here is generalizable to any domain, the specific examples explored are within the genomics domain. The final tool, GeneSet MAPR, is especially useful in a domain with little ground truth and a huge volume of noisy, uncertain data. We show that GeneSet MAPR performs better at discovering related but concealed data points than an approach using the same data without abstraction, as well as a an established state-of-the-art approach that works on a network but ignores the heterogeneous patterns. It does this while providing details the other methods cannot.

To my amazing wife and hilarious, curious sons, who treat me like the smartest person in the world. To the brilliant faculty at UIUC, who showed me otherwise. To the Curry Train, Logo Lunch, and Catering by Wahlburgers, who helped me split the difference.

ACKNOWLEDGMENTS

The method presented in this thesis, GeneSet MAPR, was created to be part of the suite of tools available on the KnowEnG platform developed here at the University of Illinois and funded by the National Institutes of Health. It would not have been possible without the generous support of both institutions.

Special thanks go out to my advisor, Shobha Vasudevan, who recognized my interest in the subject and offered the opportunity to work on the project, as well as Professor Saurabh Sinha and Charles Blatti, who helped focus and evaluate the work. Thanks are especially due to Charles, with whom I spent countless hours debating topics ranging from some of the most minute implementation details to much higher-level domain-specific context. Thanks as well to Dr Krishna R. Kalari with Mayo Clinic, who provided the data used for the application example in Chapter 6, and took the time to meet and offer feedback on our findings. There are many others associated with KnowEnG off of whom I bounced ideas, or who otherwise offered interesting directions for this research — too many to name, but I appreciated their time and enthusiasm.

Finally, my deepest gratitude is for my family. Not only did they patiently bear with me as we uprooted our lives over these past few years, but their support has been downright saintly. I am happy that after all this uncertainty our lives will soon be more stable, and I look forward to spending more time with all of them.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation

Suppose one has a dataset — say, our mother's music collection — to which we wish to contribute. First, we need to understand the collection. Which is more representative of her taste: the Phish albums, the Motown collection, or the Bing Crosby she plays for the holidays? We have imperfect metrics to define her preferences: how often she plays an album, or how much dust is on the cover. But we also have a vast wealth of outside information, details for every album including genre, artists, era, which band plays another's covers, whose band members have played together, when and which songs appear on the Billboard Hot 100 or Prairie Home Companion, and much more. Here we present an approach that leverages established domain knowledge to learn what makes a dataset unique, and applies the patterns from within that dataset to find other similar items, so one may buy Mom the perfect birthday present. Of course, this also generalizes to other applications.

The problem can be thought of as one of *set membership*, where some outliers are less representative of the set as a whole and many items which should belong in the set are yet to be discovered. Similar datasets exist in many domains: biology, movie-streaming services, social networks, book sellers, and personalized clothing. Today there is even a company whose focus is on recommending the perfect coffee! In every case, a set of provided items is deemed to be important. The first challenge, then, is in setting an objective function that captures the uniquely identifying characteristics within the set, based on a compendium of a priori domain knowledge. The second challenge is to avoid *over-fit*, where noisy, inconsequential details nonetheless show some correlation to the problem at hand, such as using the Super Bowl winner to predict the stock market [1]. An effective next step is

to arrange the domain knowledge in the form of a network (or graph) [2, 3, 4]. Networks are used in many domains to represent a collection of items which share pair-wise relationships, and readily allow for the formal quantification of higher-order relationships between items.

This problem has similarities to both *personal recommendation* and *community detection*. Unlike personal recommendation, which is often more interested in marketing a diverse array of products or media in an effort to entice a user, we wish to be more careful about Type I error, or the erroneous inclusion of items which should not belong. We also wish to return an intuition for what it is that makes the set unique. In community detection, one attempts to cluster all items in a network into semantic groups, identifying and separating unique chunks [5, 6]. Older methods tended to require the user to provide an expected number of clusters while newer methods, recognizing the problem of such an upfront requirement, attempt to use ideas such as cut-minimization, flow, or spectral partitioning to learn the optimal number of clusters in an unsupervised manner [7]. For example, depending on the fashion at time of publication, in academic literature terms such as Machine Learning and Artificial Intelligence may be used interchangeably or may instead refer to distinctly separate processes. Applying a community detection algorithm may uncover modules within that literature with distinct algorithmic approaches, such as might be seen in the fields of Signal Processing versus Robotics versus Computer Vision. While the goal of community detection is to optimally categorize all items in the network, we wish to apply a semi-supervised approach that takes a user's input and finds the items most likely to belong in the cluster represented by that set. What other clusters, or how many, may exist in the network is not our concern.

To this end, we present Geneset MAPR (Meta-path Analysis for Pattern Regression) an algorithm that learns the patterns of connectedness within a set, based upon a network containing various types of domain knowledge. MAPR then applies those patterns to give a probabilistic score to all items in the network based on how similar they are to the input set. Whereas it is common for a network-based similarity measure to use a network of homogeneous relationships or otherwise make no real distinction between them [8], MAPR leverages the ideas of *meta-paths* and *connectedness* to explicitly account for the heterogeneous nature of underlying relationships. Meta-paths allow us to apply a level of abstraction to the data — smoothing

noise in favor of broader trends — while providing a quantifiable network-based pattern from which we can create a set's identifying fingerprint. An item is deemed similar to the set if the pattern of connectedness joining it to the set is similar to the uniquely representative pattern for items already included. Meta-paths have been shown to be useful for community detection in the academic literature domain, provided a user specifies the number of clusters, a few items to seed each cluster, and a small set of interesting meta-paths [9, 10]. However, one of our primary goals is to remove from the user the responsibility for choosing appropriate meta-paths (while leaving open the possibility for a user to later weight the returned meta-paths according to her own preference). As mentioned, we also omit the requirement to pre-select a number of clusters; defining an optimal number of clusters is a very subjective problem whose answer is very dependent on context. Our goal is to return how likely an item is to belong in the user's preferred cluster. While GeneSet MAPR was developed to apply these ideas to the genomics domain, the framework is readily generalizable with little to no adaptation to existing code.

1.2 GeneSet MAPR and the Genomics Domain

Gene set analysis is a mainstay of modern functional genomics studies. Relatively fast, high-throughput experiments enable a researcher to quickly identify a set of genes based on a preferred metric, such as gene expression levels, altered epigenomic states, or signatures of selection in coding sequences. For example, differential expression may then be compared between patients who recovered after a given treatment versus those who did not in order to define a phenotype. Such measures are typically supplemented with a p-value, an estimation of how unlikely it is that an observation is due to random chance. Given such a gene set, a researcher will attempt to learn its relevant properties. This is commonly done by comparing overlap with established but continuously evolving functional annotations, such as gene ontology terms or protein domains. A term that has high overlap with the gene set is said to be *enriched* for the set. The most basic example of this would be Fisher's exact test, but tools such as DAVID and GSEA offer somewhat more advanced enrichment analysis with the advantage of convenient interfaces.

In the genomics, and the field of bioinformatics more generally, ground truth can be hard to come by. Experimental validation of a finding often requires expensive, targeted experiments with long lead times. Additionally, clinical trials often suffer from small sample sizes, and focusing on a particularly lethal or specialized variation of a disease makes it even harder to find participants. Obviously, a small sample size makes measures of magnitude, such as differential expression, more susceptible to noise in the data. Furthermore, there can at times be a misplaced reliance on statistical significance, or p-value, as a hard threshold without accounting for the statistical power of the study or other factors [11, 12]. In the worst case there may be bias towards presenting results that conform with other published research, and are therefore considered more believable.

Notably, the types of measurements that lend themselves to a high-throughput setting are often unrelated to causality. For example, a biological process often relies on a series of interactions between genes, and an upstream mutation in one gene may cause a change in the expression levels of several other genes. The downstream genes may be flagged by the chosen experiment while the upstream gene — the cause of the change — is not. Ideally, we wish to find the pattern of connectedness between genes in the user-provided set that guides a researcher to overlooked genes that nonetheless belong in the set, allowing her to spend limited resources more efficiently.

GeneSet MAPR applies a set membership approach to gene set analysis, leveraging not only multiple varied networks of gene annotations, but also many disparate forms of gene-gene relationships. This can be thought of as establishing a weighted combination not only of enriched terms, but of whole bodies of research that is uniquely descriptive of an input gene set. In the field of bioinformatics, networks exist representing relationships between patients, diseases, genes, processes, some combination of these, and more. Here, we concern ourselves specifically with networks encoding interactions and relationships at the gene level. We make no judgments as to the quality of the networks; rather, we accept them as established knowledge that will help to better understand the gene sets in question. Of course, the quality of underlying networks does matter, but as shown later in this thesis the approach used by MAPR is able to emphasize data that is relevant to a user's input set.

1.3 Outline

Before outlining and evaluating GeneSet MAPR, this thesis first reviews necessary concepts pertaining to networks in Chapter 2, culminating in the definition of meta-paths. Next, the networks upon which MAPR was tested along with their sources are detailed in Chapter 3. Chapter 4 describes the approach used by MAPR, while Chapter 5 presents an array of evaluations to validate and elucidate its performance, including comparisons with state-of-the-art approaches to this problem that have been applied in the genomics domain. A real-world application to a novel, newly defined gene set appears in Chapter 6. Finally, proposed next steps and concluding remarks follow.

CHAPTER 2

NETWORK STRUCTURE AND META-PATHS

GeneSet MAPR considers relationships between genes in a set in order to gain an understanding of what makes the set unique from the rest of the genome. Therefore, it requires some collection of accepted knowledge upon which to draw, even as we grant that such knowledge may contain imperfections. In mathematics, a structure representing a collection of items sharing pair-wise relationships is called a *graph*. In many other domains, it is frequently termed a *network*. So, too, do the names of components within such a structure differ between domains. Therefore, we will now introduce the terminology to be used for the rest of this paper as we build towards the definition of a *meta-path*, the utility of which is the hypothesis underlying GeneSet MAPR's development.

Readers interested in a deeper understanding of graph theory, including proofs and additional concepts are encouraged to see [13, 14].

2.1 Basic Network Terminology

In biology, the word *network* is often used to describe what is formally known in mathematics as a *graph*, the most basic component of which is called a *vertex*, or *node*. An example can be seen in Figure 2.1. Each node represents a discrete data point, and a network encodes the relationships between them. In the context of this thesis, nodes may represent either a specific gene or a non-gene term that describes some shared relationship among a group of genes, such as ontological annotations, protein families, or interaction pathways.

A relationship between a pair of nodes is represented by an *edge*. For example, each gene belonging to a given protein family — a non-gene term — will be connected to that family node via an edge. The *degree* of a node is

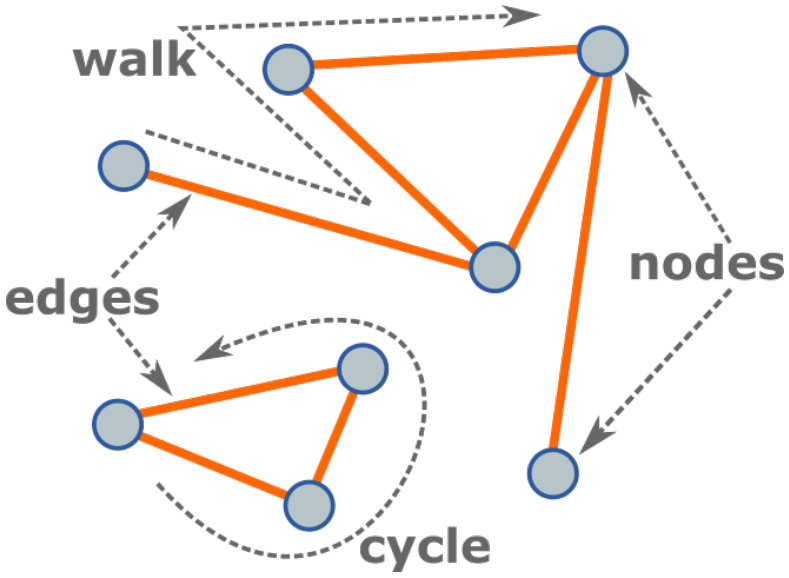


Figure 2.1: A simple network showing examples of nodes, edges, a walk, and a cycle.

the number of edges connecting to, or shared by, that node. In this case, the node for a large family may have a degree of 1,000 or so, as it will have a separate edge connecting it to each of its member genes. Meanwhile, a typical gene node will have a much lower degree, as its membership has only been confirmed for a handful of protein families, and thus it only shares a handful of edges. A small number of nodes corresponding to well-studied genes have a much higher degree, as more relationships have been experimentally discovered. Nodes with exceptionally high degree are called *hub nodes*. What constitutes a hub node is contextual, but generally it is a node whose degree far exceeds what would be expected if all the edges in a network were randomly distributed.

Edges may be *directed*, representing a one-way relationship. A relationship that is mutual, or reciprocal, is represented by an *undirected* edge. For example, if two genes in a homology network share an ancestral origin, then their corresponding nodes will be connected by an undirected edge indicating that shared, mutual relationship. In contrast, if one gene acts upon another, or if an ordered series of relationships is important, a directed edge will connect the origin gene to the terminus. As mentioned in Chapter 1, we are interested in discovering both upstream and downstream genes: those that may have caused the observed behavior as well as those that may be affected

by it. As such, the GeneSet MAPR method presented in Chapter 4 treats all relational edges as undirected.

Edges may have a *weight*. The weight may be binary, where the presence of an edge indicates merely the existence of the specified relationship. In this case, the weight is 1. Alternatively, edge weight may indicate the strength of a relationship, the frequency of its occurrence, or the confidence that a gene truly belongs in the specified protein family. In such a case, the lack of a connecting edge between two nodes has the same implication as creating an edge with a weight of 0: there is no observed relationship between those two nodes.

Many pairs of nodes may not share any edges; that is, they may not be directly connected. However, they may share an indirect, multi-step connection called a *walk*, where an originating node shares an edge with a second node, which shares an edge with a third node, and so on until reaching the specified terminal node. There may exist many unique walks connecting the two nodes, each passing through different sets of nodes and edges along the way. A *path* is similar, except that it specifies only the edges encountered at each step along the walk. Path *length* denotes the number of edges that are part of the path. A *cycle* refers to a walk or path that begins and ends at the same node. A path itself may not be a cycle while nonetheless containing one. A *simple path* denotes the absence of any such cycles: no nodes or edges are repeated.

2.2 Homogeneous and Heterogeneous Networks

Edges may have an additional property called *type*. This is used when an individual edge in the network may represent one of several disparate types of relationships. For example, two gene nodes may share several different edges, as they may be related through homology, direct protein-protein interactions (PPI), and frequent co-occurrence in published research. Separate edges would be used to indicate each relationship, with corresponding edge types of homology, direct PPI, and textmining. Each edge would be weighted according to the methodology for its respective edge type, and prior to any standardization the edge weight is relative only to other edges of the same type.

In many networks, all edges are of the same type. Such a network, termed *homogeneous*, describes a uniform collection of interactions between similar objects. A *heterogeneous* network, on the other hand, describes a network consisting of multiple types of edges and nodes. In the KnowEnG network introduced in Chapter 3, gene nodes may be interconnected via homology or interaction edges, and may also be connected to non-gene term nodes via ontology or protein family edges. Heterogeneous networks introduce a wider range of independent observations, but require different considerations than homogeneous networks.

A *subnetwork* of a given network N is the network formed when only a subset of edges and nodes from N are used. Unless otherwise specified, in this paper we will use *subnetwork* to refer to the homogeneous network induced by extracting all edges of a specific type. That is, one could think of the heterogeneous network of multiple typed edges as having been constructed from the union of several homogeneous subnetworks, as $N = \{N^{t_1} \cup N^{t_2} \cup N^{t_3} \cup \dots\}$. Any nodes connected to those extracted edges are also considered a part of the subnetwork. Thus, for our purposes, specific nodes may appear across multiple subnetworks but an individual edge will belong only to the subnetwork of corresponding edge type.

An illustration of combining two homogeneous subnetworks into a single heterogeneous network can be seen in Figure 2.2.

2.3 Meta-Paths

We are interested in evaluating structural patterns between nodes in a way that incorporates the heterogeneous nature of the relationships contained. However, we also wish to ensure the evaluations can be done with reasonable computational efficiency and robustness to noise. To this end, we explored the use of an abstraction termed *meta-paths*.

The objective in using meta-paths is to describe potential paths of varying lengths between nodes at the meta level — a higher level of abstraction than individual edges. A meta-path defines a new form of relationship between nodes of the same type as a composite of specified edge types [9]. That is, for a network comprised of a collection of multiple edge types $\{t_1, t_2, \dots\}$, a meta-path m specifies the ordered series of edge types which must be tra-

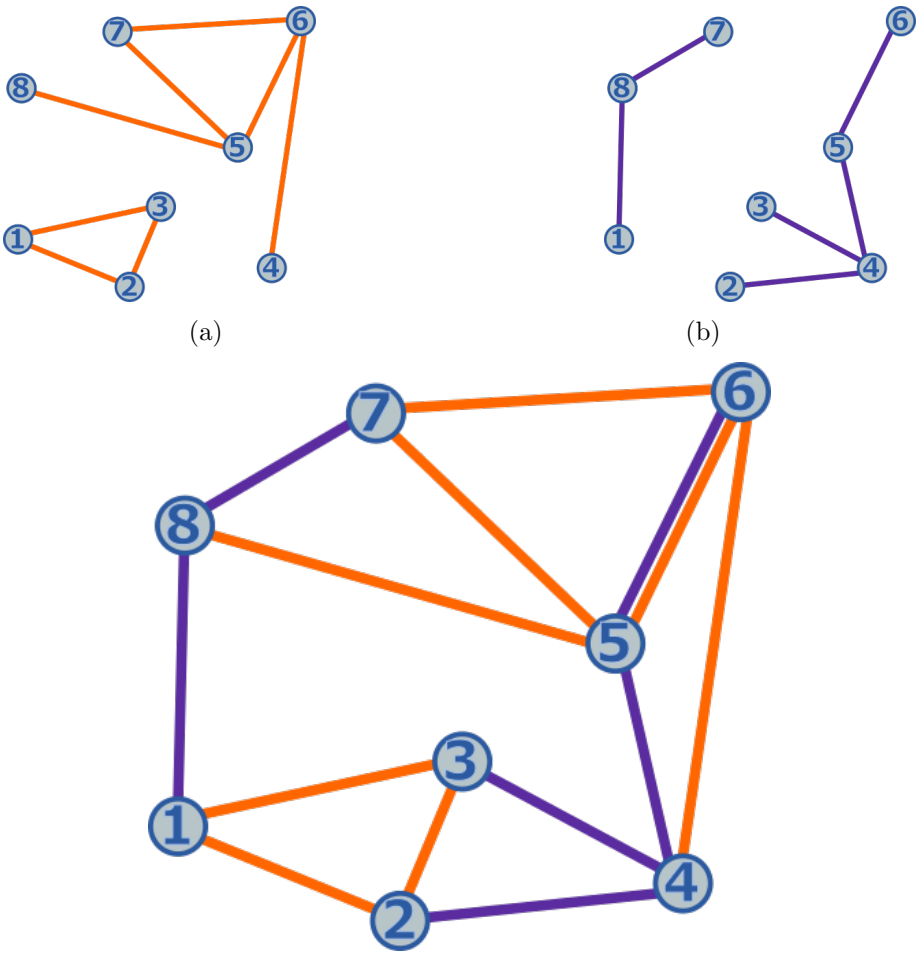


Figure 2.2: (a) and (b) show homogeneous subnetworks, each with its own set of connections over the same set of nodes. (c) shows the combined heterogeneous network with edge type, or originating subnetwork, specified by color. Supposing meta-path m_1 is defined as (**orange**, **purple**), we observe that node 6 is connected via m_1 to nodes 2, 3, 4, 5, and 8, ignoring cycles. Following the inverse meta-path $m_1^{-1} =$ (**purple**, **orange**), node 6 is connected via m_1^{-1} to nodes 7 and 8.

versed for the connecting path to be considered. For example, consider a network containing two edge types: homology (Ho) and protein interaction (Pi). Suppose we define meta-path $m_1 =$ (Ho, Pi, Ho), where the relationship described by m_1 is one where the origin and terminus genes each have a homologous gene (related through ancestry) that share a physical interaction at the protein level. Thus, our analysis of m_1 would only consider connections between genes where the edges along the path follow the pattern (Ho,

Pi, Ho). We could then specify a collection of meta-paths $M = \{m_1, m_2, \dots\}$ which are deemed of enough interest to warrant investigation. An illustration of meta-paths in a heterogeneous network can be seen in Figure 2.2.

Supposing the edges in the network are all undirected, a *symmetric* meta-path is one where the specified order of edge types is itself symmetric. Thus, if gene g_1 connects to gene g_2 along the symmetric meta-path $m_1 = (t_1, t_2, t_1)$, then the reverse is true. That is, $g_1 \xrightarrow{m_1} g_2$ implies $g_1 \xleftarrow{m_1} g_2$. However, this is not the case for an *asymmetric* meta-path, such as $m_2 = (t_1, t_2)$. So we will define an *inverse* meta-path as one where the order of edge types is reversed, such that $m_2^{-1} = (t_2, t_1)$. It follows that in the case of asymmetric meta-path m_2, $g_1 \xrightarrow{m_2} g_2$ implies $g_1 \xleftarrow{m_2^{-1}} g_2$.

In Chapter 4, we outline a method that collects and compares the number of paths connecting genes within an input set, provided those paths follow the pattern of edge types described by select meta-paths. As short-hand, we will often refer to two genes as *connecting along meta-path* m if they are connected by one or more paths of the pattern specified by either m or m^{-1}.

CHAPTER 3

THE KNOWENG NETWORK

The GeneSet MAPR method explored in Chapter 4 and the tools against which it was evaluated require networks upon which to run, to be used as compendia of a priori knowledge. For this, we use the networks collected for use by KnowEnG (Knowledge Engine for Genomics), outlined in Table 3.1. MAPR was designed to be part of the suite of tools available on the KnowEnG platform developed here at UIUC (University of Illinois at Champaign-Urbana) and funded by the NIH (National Institutes of Health). For use with its network-based analytics tools, KnowEnG has collected data from many large, publicly available sources.

For the development and evaluation of GeneSet MAPR, nine distinct subnetworks were extracted, representing a wide range of potential relationships and annotations. Some, such as Gene Ontology, are widely accepted ways of understanding studied relationships. Some, such as Textmining, are much less traditional. While the networks used by KnowEnG have undergone continued refinement and expansion over the past two years, the beta versions of the networks used to evaluate MAPR were frozen at the start of its development. This ensured fair comparisons of MAPR to previous states, as well as to the state-of-art network-based DRaWR in Chapter 5. One exception to this is the Homology edge, and the reason for the update will be addressed below.

3.1 Network Contents

3.1.1 Shared Annotations

Three of the subnetworks used in this paper represent gene *annotations*. When a group of genes have all been found to share some association — such

Table 3.1: Summary of Subnetworks

Subnetwork	Terms	Weight	Source	Description
coexpression	n	probabilistic	STRING	predicted association based on expression pattern
GO: Bio Proc	y	1 or 2	GO Cons.	larger processes composed of multiple gene products
GO: Cel Comp	y	1 or 2	GO Cons.	regions where gene products are active
GO: Mol Func	y	1 or 2	GO Cons.	molecular activities of gene products
homology	n	$-log_{10}(E)$	BLAST	shared ancestry between a pair of genes
pathway	y	binary	KEGG	series of genes contributing to cellular behavior
PPI: direct	n	binary	DIP	direct protein-protein interaction
PPI: genetic	n	binary	DIP	genetic protein-protein interaction
PPI: physical	n	binary	DIP	physical similarities between proteins
Protein Fam	y	$-log_{10}(E)$	Pfam	protein domain: indicates relationship through evolution
textmining	n	probabilistic	STRING	statistically relevant co-occurrences in scientific texts

as all playing a significant role in brain development, or interacting in the same signaling pathway — the genes in that group are all labeled with that specified annotation. Individual genes typically share many annotations, which can vary greatly by size and the amount by which the annotations overlap. In these respective subnetworks, the annotations are represented as non-gene nodes, which will also be referred to as *term nodes*.

In this version of the KnowEnG network, each of the annotation subnetworks is *bipartite*. In a bipartite network, the nodes can be separated into two groups, where edges connect nodes from one group to another, but do not connect nodes within either group. So in these subnetworks there are gene nodes and term nodes. Every edge in the network connects to both a gene node and a term node, never joining two nodes of the same type. Term nodes may have a degree anywhere between a few and several thousand edges, representing all of the genes belonging to that annotation term.

Gene Ontology

The Gene Ontology (GO) project set out to define an ontology — a representation of collected knowledge — regarding genes and their functions. The explicit goal was to provide a structured and controlled vocabulary for use in gene annotation that was biologically meaningful and unified across species [15]. Ontology terms and definitions are made available through the GO Consortium.

In the KnowEnG network, a GO edge has two possible edge weights. A value of 2 indicates an experimentally supported annotation, where the association between a gene and the function described by a GO term is shown through direct experimentation and supported through published, peer-reviewed research. A value of 1 indicates an inferred relationship, where a gene's GO annotation is inferred from the annotation of other genes within the same family, or with a shared biological ancestry.

It should be noted that the functions described by GO terms can be related to each other, resulting in a hierarchical network. At the top are three terms, each describing a different potential aspect of any given function. Terms falling under **molecular function** describe activities at the molecular level. **Cellular component** terms refer to the location within cellular anatomy where a molecular function is carried out. A **biological processes** term describes a multi-step series of events involving one or more molecular functions. This is not the same as a pathway, which is discussed next. In conversations with biologists, it was suggested that these three aspects are distinct enough that we may wish to separate the GO network into three respective subnetworks, based on these top-level terms. Indeed, this change brought two benefits: providing more intuitive edge types for describing how sets are connected, and immediately improving GeneSet MAPR's ability to predict hidden members of a set, an evaluation metric explored in Chapter 5. The GO subnetwork used for this paper did not incorporate relational edges between GO terms, thus the hierarchical nature of the GO network was not addressed by MAPR. Chapter 7 discusses how this could potentially be incorporated.

Pathway, KEGG

A pathway describes a series of inter-dependent actions effecting some change in or product produced by a cell. For example, activating a pathway may result in turning a gene off, or assembling a new protein. Common pathways relate to metabolism, gene regulation, and signaling. A pathway term node shares edges with all genes involved in that series of interactions. Edges are binary, indicating that each specified gene has been implicated in that respective pathway.

KnowEnG collected terms in this subnetwork from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16], a pathway database established by Kyoto University. Each manually drawn pathway relates molecular-level information to higher-level functions and utilities through genes and their products.

Protein Domain, Pfam

A domain specifies a distinct region on a protein sequence that is stable and can fold independently of the rest of the protein. A specific domain may appear across many proteins, in varying combinations with other domains. Edges in the subnetwork indicate an association between a given gene and family of protein domains. Edge weights are continuous values based on the Expect value, which describes the expected number of matches found by making a similar random selection from a database of the same size. The closer to zero the Expect value, the more significant the match. Final edge weights are the negative log 10 of the Expect value.

This subnetwork comes from Pfam, produced at the European Bioinformatics Institute [17]. Pfam uses hidden Markov models to identify and align protein domain family members.

3.1.2 Direct Relationships

The remaining subnetworks each represent a specific type of direct gene-gene relationship. That is, the nodes in each represent individual genes and nothing else. Edges exist between a pair of genes only if the specified interaction between the two has been observed, or otherwise inferred.

Homology, BLAST

Homology refers to the existence of a shared ancestry between two genes, which have since evolved independently. Similar to the protein domain subnetwork, edge weights are negative log 10 of the Expect value. Edges were collected from Protein BLAST (Basic Local Alignment Search Tool), or Blastp, created by the U.S. National Library of Medicine [18].

The homology subnetwork was the only one to be updated over the course of GeneSet MAPR's development. We noticed during evaluations that the homology subnetwork consistently showed little to no positive impact on results across all tested gene sets, as compared to other subnetworks. Further examination showed it was surprisingly sparse — containing many fewer edges than would have been expected — suggesting a poorly defined threshold or other issue when the subnetwork was collected. The KnowEnG network had undergone many changes in the meantime, so we swapped in the Blastp homology subnetwork and re-ran the evaluations, noting an improvement in the results.

Protein-Protein Interactions

Processes within a cell often result from the interaction of many proteins. Such a Protein-Protein Interaction (PPI) generally designates a chemical reaction catalyzed by two or more proteins in close physical proximity. There are different categories of PPI, and they may be measured or deduced using different techniques. Here, three types of PPI are considered, each of which is represented by its own subnetwork: **direct interaction**, **genetic interaction**, and **physical association**. In each case, edge weights are binary, representing merely the existence of a relationship.

All three networks were collected from the Database of Interacting Proteins (DIP), compiled by the University of California, Los Angeles [19].

Textmining, STRING

Textmining represents a much different approach to automatically inferring and collecting gene-gene interactions. Here, an automated process scans Medline abstracts as well as a significant number of full-text articles, uncov-

ering both semantic connections between genes and statistically significant shared occurrences [20]. Edge weights are based on a probabilistic score combining evidence from multiple channels while adjusting for the likelihood of randomly observed interactions.

KnowEnG collected textmining edges from the STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins), a collection of known and predicted functional interactions [20].

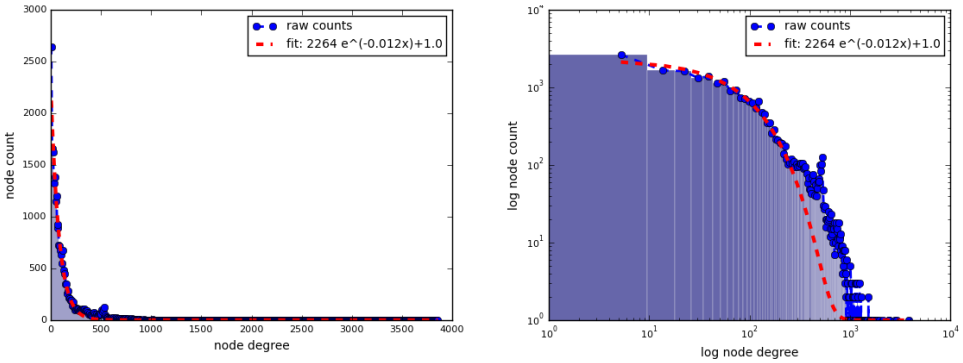
Coexpression, STRING

The coexpression subnetwork is perhaps the most unconventional network used in this thesis. Here, shared edges represent gene-gene interactions predicted by algorithmic analysis of genomic information, as well as analysis of simultaneous gene expression (hence, coexpression) [20]. As STRING is the source of this subnetwork, edge weights are based on the same probabilistic scoring method as the textmining subnetwork.

3.2 Distribution of Node Degree in Network

A *scale-free* network is one wherein, with respect to node degree, there exists no "typical" node according to which the rest can be characterized [21]. There is such variance in the number of edges connecting to individual nodes that the network lacks a relative scale. This means there is an exceptionally high number of nodes with very few connections, yet also a high occurrence of nodes with a degree far greater than the average. Scale-free networks are common in many domains, and biology is no exception [22]. In fact, the KnowEnG network used in this thesis follows a similar distribution. The median node degree in the network is 69, while the mean is nearly twice that at 127.7.

In many ways, this imbalance is to be expected. Some genes and processes have a much longer history of focused study than others, and that history can in turn influence where a researcher chooses to spend her limited resources, compounding the issue. The result is that some genes have a wealth of established, experimentally proven connections while there is less incentive to study those that are much less established.



(a) (b)

Figure 3.1: Histogram of nodes in KnowEnG network by node degree. Right is a log-log transformation of the left figure. The dotted line shows an exponential function fitted to the raw counts of nodes.

Figure 3.1(a) shows the histogram of KnowEnG network nodes by degree, and an exponential function that has been fitted to the distribution. Indeed, the distribution of node degree approximates the power law distribution, more easily seen in the log-log scaling of Figure 3.1(b). Such networks can prove challenging to work with, as hub nodes may wield undue influence, and divining useful information from a collection of very sparsely connected nodes may prove difficult. One goal of GeneSet MAPR is to put these genes on a more equal footing, such that a researcher may be guided to invest time in — and discover fruitful connections to — genes which are currently less-understood. The approach taken by MAPR to address this is outlined in Chapter 4, and an evaluation of the final result appears in Chapter 5.

CHAPTER 4

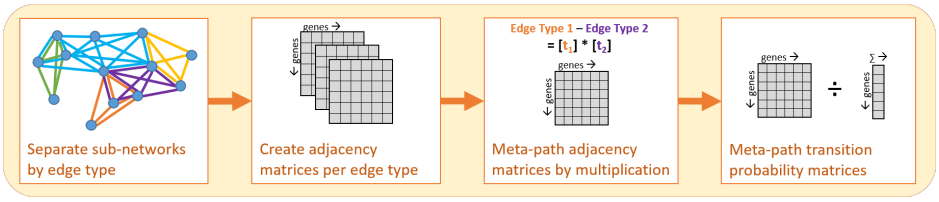
GENESET MAPR

Here, the approach used by GeneSet MAPR is described in detail. The goal of MAPR is two-fold: to learn the pattern of relationships connecting genes within a set, and to use that learned pattern to rank all available genes by their similarity to the set. The process by which this is done revolves around quantifying the connections along meta-paths in a heterogeneous network of a priori knowledge.

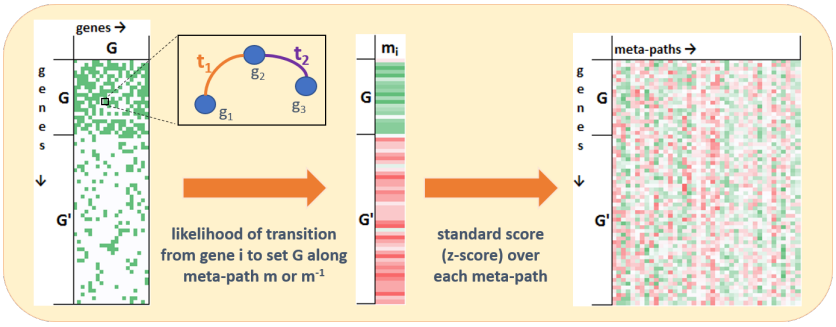
To this end, GeneSet MAPR performs three primary steps. First, the heterogeneous network is pre-processed to speed up the retrieval of meta-path counts. Second, meta-paths are used to define an abstracted pattern of connectedness within a gene set, to be used as set-specific features. Third, genes are ranked by their connectedness to the set using an ensemble of LASSO regressions.

There are a few challenges posed by the genomics domain to any classification or similarity algorithm that GeneSet MAPR must overcome. Robustness is one. For instance, the observations in the underlying networks will be subject to some noise: errors in measurements as well as, in some cases, a shifting understanding of gene-gene interactions and their import. Additionally, some genes and their products have been the focus of much more study than others, resulting in unbalanced subnetworks, where a few hub nodes have hundreds or thousands of connecting edges while most nodes have less than 50. By viewing the network in terms of meta-paths, MAPR introduces a level of abstraction that helps smooth noise at the level of individual edges. As for hub nodes, their out-sized influence is mitigated through both the gene-relative transition matrices (Section 4.2) and path-relative connectedness score (Section 4.3); a quantitative examination regarding influence of node degree is provided in Chapter 5.

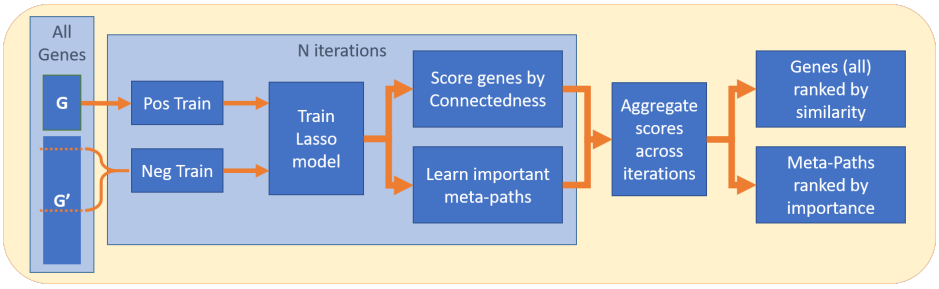
The gene sets themselves may be noisy as well: they are often created by applying a threshold to one or more measurements, whose values may be



(a) convert meta-paths to transition matricies



(b) calculate connectedness to gene set



(c) train ensemble of classifier models

Figure 4.1: Overview of GeneSet MAPR

derived from a small number of study participants. Thus, there is uncertainty in which genes are deemed part of the positive set. Furthermore, unlike classical machine learning formulations, there is no representative negative set. What GeneSet MAPR is presented with, then, is a positive set — the input gene set — typically of around 50–500 genes. In the case of the human genome, this is out of 23,782 genes with connections in the KnowEnG network, the rest of which are treated as unlabeled. MAPR approaches these challenges by training an ensemble of classifiers, and sub-sampling the unlabeled and positive sets as necessary (Section 4.4).

A visual reference of MAPR's primary steps can be found in Figure 4.1.

4.1 Conversion of Indirect Connections

Table 4.1: Edge Counts Before/After Conversion

subnetwork	original edges	term nodes	converted edges	edge type in network
coexpression	–	–	50,100	STRING_coexpression
GO: Bio Proc	142,699	10,905	16,773,233	GO_BioProc
GO: Cel Comp	71,627	1,476	73,632,134	GO_CelComp
GO: Mol Func	62,624	3,815	48,104,940	GO_MolFunc
homology	–	–	652,823	blastp_homology
pathway	24,878	299	2,292,622	kegg_pathway
PPI: direct	–	–	86,569	PPI_direct_interaction
PPI: genetic	–	–	2,487	PPI_genetic_interaction
PPI: physical	–	–	184,435	PPI_physical_association
Protein Fam	71,324	3,631	5,423,388	pfam_domain
textmining	–	–	354,931	STRING_textmining
total	373,152	20,126	147,557,662	

As discussed in Chapter 3, the network of a priori data used during evaluation of GeneSet MAPR consists of 11 subnetworks, each describing a different relationship type or observation methodology. Five of those subnetworks consist of shared annotations, rather than direct gene-gene relationships. In this section, such subnetworks will be referred to as containing *indirect* connections, where a gene-gene connection must pass through an annotation, or a term node.

Subnetworks consisting of indirect connections pose a philosophical conundrum. We initially wish to remain agnostic to the types of underlying subnetworks, giving equal weighting to each, while leaving open the possibility for the end-user to later specify the importance they wish to apply to a given subnetwork. In the case of meta-paths, an indirect gene-gene connection must be represented by two steps along the same edge type, as opposed to just one for a direct connection. However, since these indirect subnetworks are all bipartite, they can easily be reduced to direct gene-gene connections.

For a bipartite subnetwork of indirect connections, GeneSet MAPR converts each annotation term along with all of its shared edges into a corresponding set of gene-gene edges. If two genes each share an edge with the same annotation, those two edges will be replaced with a single edge between those two genes. Edge weights are normalized per subnetwork to the range $[0, 1]$ and treated as the confidence in the existence of a particular edge — a probabilistic estimate that the annotation or relationship is accurate. The

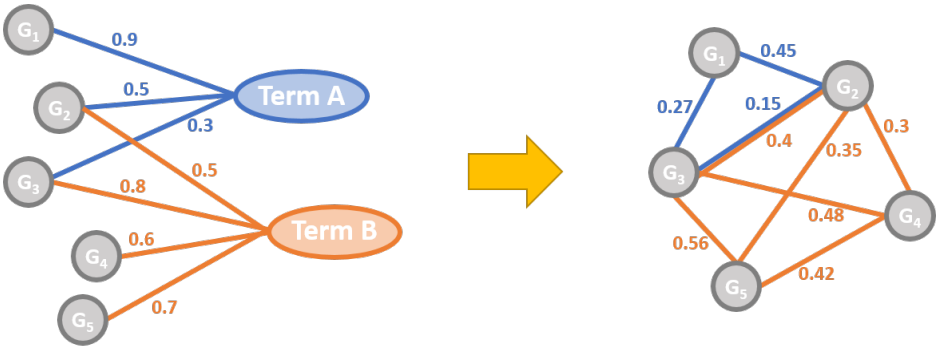


Figure 4.2: Simple example of converting indirect gene-term edges into direct gene-gene edges. Left is the original subnetwork with two annotation terms. Right is the final result, where edge weights between two genes are the product of the two indirect edges connecting them.

weight of the new edge is the product of the weights of the two replaced edges. Hence, each gene then shares a direct connection to every other gene annotated by the same term, where the new edge weight is the joint probability that the two genes share that annotation. If two genes share multiple annotations, then one new edge will be created between them for each annotation. An example of this conversion can be seen in Figure 4.2.

Table 4.1 shows each subnetwork used in this thesis, along with a corresponding count of edges contained. For subnetworks containing indirect connections, the count of term nodes and original gene-term edges is given. The converted edges column shows the total count of direct connections after converting any term nodes to gene-gene edges. In each case where indirect edges were removed, the resulting converted subnetwork was considerably more dense.

4.2 Quantifying Meta-Paths

One challenge lies in quantifying the meta-paths within a heterogeneous network while giving equal emphasis to each homogeneous subnetwork. Subnetworks may be sparse or dense and may contain multiple edges between gene pairs, and edges may be weighted to indicate confidence or else simply represent the binary presence of a relationship. GeneSet MAPR approaches this issue by describing each subnetwork as a normalized adjacency matrix, and each meta-path as a probabilistic transition matrix.

Adjacency matrix An adjacency matrix is defined as a square matrix describing the potential connections between every possible pair of nodes in a network [14]. MAPR creates an adjacency matrix A^t for every subnetwork, specified as the homogeneous network containing all edges of type t. Traditionally, an entry will have a value of 1 if node i shares a connecting edge e with node j, and 0 otherwise. MAPR modifies this definition (4.1) to account for networks where genes may share multiple edges, and edges may have weights indicating the strength or confidence of that relationship. Values are normalized to the range $[0, 1]$ to facilitate comparison between disparate network types (4.2). Additionally, each edge is treated as undirected, where the presence of a relationship from i to j implies a mutual corresponding relationship from j to i. Thus an adjacency matrix for a given edge type is symmetric.

$$\widetilde{A}^t \leftarrow \tilde{a}^t_{ij} = \sum e^t_{ij} \tag{4.1}$$

$$A^t = \frac{\widetilde{A}^t}{max(\widetilde{A}^t)} \tag{4.2}$$

Transition matrix If one considers starting at a given node and taking a single step along any of its connecting edges, then a transition matrix describes the probability of transitioning from any row node i to any column node j. Each entry is a non-negative value, and every row and/or column must sum to 1 [23].

MAPR first finds a meta-path adjacency matrix $\widetilde{B}^m$ by multiplying normalized edge adjacency matrices A^t in the order specified by the meta-path m (4.3), such as $m = (t_1, t_2, t_3)$. Each entry in $\widetilde{B}^m$ describes the number of paths from node i to j that follow the order of edge types specified by meta-path m. Prior to each matrix multiplication, values in the main diagonals are set to 0, so as to omit the influence of loops, where a path encounters a given node more than once. MAPR avoids the consideration of loops because any connection from node i to j containing a loop may be more succinctly

described by the shorter path omitting said loop.

$$\widetilde{B}^m \leftarrow \tilde{b}_{ij}^m = \begin{cases} a_{ij}^{m[0]} & \text{if } i \neq j,\ \text{len}(m) = 1 \\ \tilde{b}_i^{(m[0],\ldots,m[n-1])} \cdot a_j^{m[n]} & \text{if } i \neq j,\ \text{len}(m) > 1 \\ 0 & \text{if } i = j,\ \forall\ \text{len}(m) \end{cases} \tag{4.3}$$

A final meta-path transition matrix B^m is calculated from the meta-path adjacency matrix $\widetilde{B}^m$ by converting each row into a probabilistic value (4.4). Note that the transition matrix is directed and asymmetric: it indicates the probability of transitioning from node i to j along meta-path m, where m specifies an ordered list of edge types.

$$B^m \leftarrow b_{ij}^m = \frac{\tilde{b}_{ij}^m}{\sum_k \tilde{b}_{ik}^m} \tag{4.4}$$

Thus, for each meta-path one wishes to consider, a pre-computed matrix is stored containing the likelihood of transitioning along said meta-path from node i to node j, where each node represents a single gene.

4.3 Gene Set Connectedness

Gene sets are often presented as complete, unambiguous sets, where membership of individual genes is binary, as opposed to probabilistic or confidence-based. Therefore GeneSet MAPR seeks to rank individual genes' similarity to the set as a whole. That is, pairwise gene-gene transition probabilities are aggregated over the input set in order to create a comparable measure of connectedness. (Section 4.4 introduces consideration for set membership ambiguity.)

Connectedness The MAPR concept of connectedness is defined along a given meta-path by taking the union of the likelihood of a gene connecting to the input set along either the given meta-path m or its inverse m^{-1} (4.5). Note that each edge in the network is considered to represent a mutual, shared relationship, and is therefore undirected. So too are meta-paths in this context considered to be undirected, where $m = (t_1, t_2)$ is the same as $m^{-1} = (t_2, t_1)$. A connecting path from the node representing gene g_2 to g_1

along m^{-1} merely represents the reciprocal relationship to that connecting g_1 to g_2 along m.

$$\widetilde{C}^m \leftarrow \tilde{c}_g^m = \begin{cases} \sum_{i \in G} b_{gi}^m + \sum_{i \in G} b_{ig}^m & \text{if } m^{-1} \neq m \\ \quad - \sum_{i \in G} b_{ig}^m \sum_{i \in G} b_{gi}^m & \\ \sum_{i \in S} b_{gi}^m & \text{if } m^{-1} = m \end{cases} \tag{4.5}$$

The standard score, also known as the z-score, describes for a normal distribution how many standard deviations σ a particular value is above or below the mean μ [24]. MAPR takes the standard score over the vector $\widetilde{C}^m$, adjusting each raw value $\tilde{c}_g^m$ into a standardized value c_g^m. The result is a value describing, relative to the typical number of connections along a specific meta-path, how much more or less connected each gene is to the input set. These column vectors are horizontally concatenated into a feature matrix X, where each row describes an individual gene's relative pattern of connectedness to the input set (4.7).

$$C^m \leftarrow c_g^m = \frac{\tilde{c}_g^m - \mu_{\tilde{c}}^m}{\sigma_{\tilde{c}}^m} \tag{4.6}$$

$$X = [C^{m_0}, C^{m_1}, \ldots, C^{m_n}] \tag{4.7}$$

The final result is a set of features that consider the heterogeneous nature of the network at an abstract level. The abstraction helps reduce some of the noise inherent in large, curated networks. MAPR is then able to compare the pattern of connectedness for individual genes against those already in the input gene set.

4.4 Gene Similarity Ranking

The *gene similarity ranking* provided by GeneSet MAPR must be robust: there may be uncertainty in the input set, and typically around 50-100 times as many genes excluded as included. Additionally, a primary goal is to move past the purely statistical view of the gene set — to identify genes related through their interactions to the set that may have failed to register a notable measured value in a given experiment. After defining the connectedness fea-

ture, MAPR's approach is to learn what uniquely defines the set by applying an ensemble of regression-based models combined with random sampling.

Ensemble Learning The goal of ensemble learning is to test multiple disparate hypotheses through the introduction of independent classifiers, or models, each of which casts a vote based on its own independently learned prediction [25]. Provided each model performs at least slightly better than random, then the ensemble, utilizing an appropriate mechanism to aggregate each model's vote, should perform better than any typical single model.

GeneSet MAPR trains an ensemble of models to overcome the noise inherent in the input set and the imbalance of the set size versus the number of unlabeled genes. A supposition here is that there exist many clusters within the genome with their own patterns of connectedness, and therefore a single model may work well on one cluster but poorly on others. The number of models MAPR trains can be set by the user; each additional regression-based model contributes to the stability of the final gene similarity ranking while extending the time required by a second or two. Unless otherwise specified, the number of models in the ensemble was fixed at 31 for the experiments in this paper.

Each model requires a positive and negative training set, along with a corresponding label. Labels are defined by an indicator function I, where a gene g is labeled with the value 1 if it is in the input set G, and 0 otherwise (4.8). The complete input set G is introduced as the positive training set $x^{\{+\}}$ (4.9). A negative training set $x_i^{\{-\}}$ of equal size is created by randomly sampling from the remaining, unlabeled genes (4.10). The vector of training labels $Y^{\{i\}}$ is provided by the indicator function (4.11).

$$I_{g \in G} = \begin{cases} 1 & \text{if } g \in G \\ 0 & \text{else} \end{cases} \tag{4.8}$$

$$x^{\{+\}} = \{g,\, \forall g \text{ if } I_{g \in G} = 1\} \tag{4.9}$$

$$x_i^{\{-\}} = \text{rand}\left(\, \{g,\, \forall g \text{ if } I_{g \in G} = 0\}\, \right),\ |x_i^{\{-\}}| = |x^{\{+\}}| \tag{4.10}$$

$$y_g^{\{i\}} = (I_{g \in G},\, \forall g \in x^{\{+\}} \cup x_i^{\{-\}}) \tag{4.11}$$

If, however, the trained model performs no better than random on the task of differentiating the positive and negative training sets, then $x^{\{+\}}$ is

randomly sub-sampled, keeping 75% of the genes in the current set, and a new $x_i^{\{-\}}$ is created. For a single model this may be repeated several times, until a subset is found that is separable by the model. For example, this may occur when an input set's overall pattern of connectedness is not uniquely discernible from that of the network of a whole, but which may contain two or more clusters with their own unique profiles. In practice, this has occurred rarely. The measure used to judge the model's performance is discussed below.

LASSO Least Absolute Shrinkage and Selection Operator (LASSO) can be interpreted as a linear regression model where values in the feature coefficients vector $\beta^{\{i\}}$ are subject to thresholding of the L1 norm (4.12) [26]. This regularization, adjusted according to the parameter λ, simplifies the learned model by forcing statistically redundant features to be given a weight of zero. Coefficients learned from the training data are then used to create predicted labels $\widehat{Y}^{\{i\}}$ for every gene in the network (4.13). Note that in this case each label is a continuous value akin to similarity rather than binary class prediction.

$$\beta^{\{i\}} = \arg\min_{\beta} \frac{1}{2N} \left\| Y^{\{i\}} - X^{\{i\}}\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1 \tag{4.12}$$

$$\widehat{Y}^{\{i\}} = \left[\hat{y}_{g_0}^{\{i\}}, \hat{y}_{g_1}^{\{i\}}, \hat{y}_{g_2}^{\{i\}}, \dots \right]^T = X\beta^{\{i\}} \tag{4.13}$$

For each LASSO model in the ensemble, GeneSet MAPR assigns a model performance score indicating how closely the predicted labels output by the model match those that were used to train it. That is, how well can the model separate its own training data. For this, MAPR uses the coefficient of determination, or r^2 [27]. In the numerator, r^2 compares the training label y_g for each gene to its predicted label $\hat{y}_g^{\{i\}}$. In the denominator, the training label is compared to the mean of the training labels $\mu_Y^{\{i\}}$. Thus, if every item is labeled correctly by the model, the fraction will approach 0 and the r^2 score will approach 1. An r^2 score of 0 indicates the model performed no better than merely assigning every item the mean value, and a negative score indicates the model failed. A score of 0 or less results in attempting a new

model based on a random sub-sample of the positive training set.

$$r^{2,\{i\}} = 1 - \frac{\sum \left(y_g - \hat{y}_g^{\{i\}}\right)^2}{\sum \left(y_g - \mu_Y^{\{i\}}\right)^2}, \quad r^{2,\{i\}} \in (-\infty, 1.0] \tag{4.14}$$

Once all the LASSO models have been trained, their predicted labels are collected. As there is no guarantee that the label values for one model are in a similar range as another, the values are standardized prior to being aggregated into a final gene similarity score s_g (4.15). This emphasizes the relative value of a label according to each model. Labels are weighted by the respective model's r^2 score, such that more weight is given to models which better separate the training data.

$$S \leftarrow s_g = \frac{1}{\sum_i r^{2,\{i\}}} \sum_i \left(r^{2,\{i\}} \frac{\left(\hat{y}_g^{\{i\}} - \mu_{\hat{y}}^{\{i\}}\right)}{\sigma_{\hat{y}}^{\{i\}}} \right) \tag{4.15}$$

Similarly, feature coefficients are aggregated from each model. In this case, the sign of the coefficient is important. The sign indicates whether the set was positively or negatively enriched for a given meta-path, that is, whether or not the input set had more connections via that meta-path than did a random sampling of the unlabeled genes. For example, it is more intuitive to show that set G displayed unusually high connectedness along m_1, as compared to other random sets, than it is to show that connectedness along m_1 was greater than along m_2, even though both may be less pronounced within G than for other random sets. To preserve the sign, feature coefficients are normalized according to the absolute maximum for that model prior to weighting by r^2 (4.16).

$$\beta_f = \frac{1}{\sum_i r^{2,\{i\}}} \sum_i \left(r^{2,\{i\}} \frac{\beta_f^{\{i\}}}{\max(|\beta_f^{\{i\}}|)} \right) \tag{4.16}$$

Thus, GeneSet MAPR returns two ranked lists in relation to the input gene set. The gene similarity rank orders all genes in the underlying network by their connectedness to the set. The *feature importance list* orders the meta-paths by how useful they were in uniquely describing the set.

4.5 Implementation Details

Several implementation details are worth noting. Note that computing every meta-path of length L or shorter in a network containing G gene nodes and N unique subnetworks will require computation time and storage on the order of $O(N^L \cdot G^2)$. This suggests a trade-off between the granularity of more subnetworks and the maximum depth of considered meta-paths. In the case of the bioinformatics domain, we chose to limit the meta-paths to a maximum length of 3 edges. The primary motivation for this decision was the desire for interpretable features: it was felt there would be little intuition for connecting paths of length 4 or greater. This consideration may change for another domain.

Additionally, as discussed in Section 4.3, not all meta-paths need be calculated. As the edges are undirected, a connection from node i to j along meta-path (t_1, t_2, t_3) is the same as from node j to i along meta-path (t_3, t_2, t_1). Omitting these inverse meta-paths reduces the number of computed meta-paths by 43% (from 1,331 to 803) for the network used here, consisting of 11 edge types and paths up to length 3. This savings asymptotically approaches 50% as N and/or L increase.

Another observation was that the raw count of meta-path connections was less important than the relative counts, as compared from one node to another. Hence, what matters in the adjacency matrices are the values relative to each other, as opposed to the scale of the values. So while the raw counts are calculated in memory as a floating point value, the matrices are scaled and converted to 16-bit unsigned integers prior to being stored. The result is a reduction in required hard drive space with no noticeable change in the algorithm's performance. There are likely many opportunities for compression and optimization that have not yet been explored.

It should be noted that, for a related reason, the normalization applied in Section 4.1 was removed from GeneSet MAPR's final implementation. In practice, the conversion of indirect to direct connections occurs concurrently with the construction of the adjacency matrix in Section 4.2. If two genes share multiple edges — multiple shared annotations — then the weights of these edges are summed during matrix creation, and the final matrix is normalized. As this is merely a matter of applying division to scale edge weights, normalizing twice is redundant. The resulting adjacency matrix no

longer represents a raw count of connecting edges, but rather the overall strength of all connections within a subnetwork between each pair of nodes.

Other considerations for reducing computational burdens appear in Chapter 7.

4.6 Source Code

All code used to run GeneSet MAPR as well as create the figures in this thesis are available in the GitHub repository glinkowski/GeneSet_MAPR (URL below). Example input and output files are also provided.

https://github.com/glinkowski/GeneSet_MAPR

CHAPTER 5

EVALUATION OF GENESET MAPR

Given a new tool, we must show how it performs. To that end, GeneSet MAPR was tested over a wide collection of gene sets against other state-of-the-art tools and methods. Several aspects of MAPR were evaluated, including how well it extracts uniquely identifying information for a gene set, its ability to aid in understanding enrichment, the effect of meta-path analysis on these tasks, and to what extent it is affected by biases in the network.

To address the first point — performance of GeneSet MAPR's Gene Similarity Ranking — we measured set membership prediction using area under the curve (Section 5.1) against another state-of-the-art method, (Section 5.4). The influence of meta-paths — implicit in comparisons against DRaWR — is made explicit in Section 5.3. The ability of MAPR to overcome network quirks is addressed in Sections 5.8 and 5.5. Other evaluations of interest also appear in this chapter.

5.1 Evaluation Metric: Set Membership Prediction

Evaluating the Gene Similarity Ranking output by GeneSet MAPR requires two things: a large number of gene sets, and an evaluation metric. The gene sets collected and tested over the course of MAPR's development are outlined in Section 5.2. Of course, there is no ground truth in measuring how similar one gene is to another; this is clear just from observing the wide variety of approaches used by the subnetworks in Chapter 3. However, we hypothesize that if a method is able to successfully identify characteristics unique to a set, then if some members of the set are hidden during training, they should still be recognized as highly similar.

The first step is to apply *cross-validation*. In all of our comparisons of

gene set similarity, we apply four-fold cross-validation, where one quarter of the gene set is concealed at a time. That is, the input set is divided into four equal sections. For each run of the method, three folds are combined and used as the input set, while the fourth fold is concealed and treated as unlabeled, along with the rest of the genome. The final version of GeneSet MAPR allows an end-user to specify the desired number of folds.

Each run of a method then outputs a list of all genes ranked by similarity. In this list, we compare how highly genes in the hidden fold were ranked, relative to the rest of the unlabeled data. We measure these results using *Area Under the Receiver-Operating Characteristic Curve* (AUC), which compares the true positive rate to the false positive rate [28]. That is, the AUC measures the area under a curve plotted in a 1×1 square where the y-axis shows how many true members of the hidden set were found and the x-axis shows how many genes outside the hidden set were falsely labeled as belonging to it. A line following the main diagonal from $(0, 0)$ to $(1, 1)$ represents the result one would expect from random selection. That is, an AUC less than 0.5 indicates the method performs no better than random. And AUC of 1.0 indicates perfect performance, where all hidden genes were labeled as more similar than any of the remaining unlabeled genes. The AUC reported for a single gene set is the mean over the four folds. While there is an argument to be made for emphasizing the Precision-Recall (PR) curve over ROC, ROC is used in many of the studies cited in this thesis and was kept for comparability. We hope in the future to incorporate the PR curve into our evaluations.

As shorthand, this process of predicting the relative similarity of genes removed from the original input set will be called *set membership prediction*. Two assumptions underlie this task: (1) There exists some set of underlying characteristics than can uniquely describe how members of the input set are interrelated. (2) The a priori data in the collected subnetworks is able to represent those characteristics.

5.2 Input Gene Sets: Collection Details

As mentioned, one of the primary evaluations of GeneSet MAPR involved comparing set membership prediction against other methods. For this we

Table 5.1: Gene Set Collections

Collection	Number of Sets	Min Size	Max Size	Source	
MSigDB	53	120	1,513	Molecular Signatures Database	[29]
dbGaP	51	69	447	database of Genotypes and Phenotypes	[30]
Achilles	75	76	541	Achilles Genetic Fitness	[31]
Allen Brain	40	290	336	Allen Brain Atlas Signature	[32]
Enrichr Path	40	101	802	Enrichr Pathway	[33]
Enrichr Pheno	40	100	1,706	Enrichr Phenotype Signature	[33]
ESCAPE	40	120	1,925	Embryonic Stem Cell Atlas from Pluripotency Evidence	[34]
GeneSigDB	40	101	1,606	Gene Signature Database	[35]
GEO	40	141	1,352	Gene Expression Omnibus	[36]
LINCS DN	40	100	228	Library of Integrated Network-based Cellular Signatures	[37]
Pathcom	40	100	1,493	Pathway Commons	[38]
Reactome	40	101	804	Reactome Pathway Knowledgebase	[39]
TargetScan	40	101	518	TargetScanHuman	[40]
GO (test)	40	103	1,291	Gene Ontology annotations	[15]

assembled several collections of gene sets, each emphasizing a slightly different focus in content and/or curation methodology. Collection sizes and data sources can be found in Table 5.1. Three benchmark collections consisting overall of 179 individual gene sets were carefully assembled from established publicly available sources. These were used to test GeneSet MAPR over the course of development and then to evaluate its performance against other methods. Later, to ensure MAPR was not over-fitting the benchmark collections, a further 400 gene sets were compiled into 10 additional collections. Genes in the evaluation sets are not assigned p-values or any other measure of statistical significance or confidence; gene set membership is provided as a binary attribute.

Many of the higher-level evaluations were performed across both the benchmark and supporting sets, ensuring the trends witnessed were not limited to a small set of samples. As will be seen, trends occurring over the three benchmark collections are also duplicated across these additional collections. This gives us confidence that the three benchmark collections are generally representative of the performance of GeneSet MAPR as a whole. Thus, some of the more time-consuming in-depth analyses are performed only over the benchmark sets.

A complete list of all gene sets can be found in Table A.1.

5.2.1 Three Benchmark Collections

MSigDB

The Molecular Signatures Database (MSigDB) is compiled by the Broad Institute and is composed of several groups. From the chemical and genetic perturbations subgroup of MSigDB's C2 collection, a group of gene sets curated from databases, literature, and domain experts [29], 53 cancer-related gene sets were selected for evaluation of GeneSet MAPR. The sets were chosen to represent a range of cancer types and contributors. Each set was identified by a given lab for a specific cancer variant and is composed of genes whose expression levels showed significant change from one set of patients to another, such as healthy tissue vs. sick, or treatment responders vs. non-responders. Each set is comprised of both up- and down-regulated genes: genes for which the expression level crossed a thresholded difference relative to the control group. For the evaluations of GeneSet MAPR, the corresponding UP and DN segments are combined into a single set.

dbGaP

The Database of Genotypes and Phenotypes (dbGaP) is a repository sponsored by the National Institutes of Health (NIH). Gene sets relate to various genetic and phenotypic datasets [30]. After discarding any containing 60 or fewer genes, 51 publicly available gene sets were randomly selected. The selected sets represent a mix of disease-related phenotypes and other naturally occurring biological characteristics.

Achilles

Project Achilles, run by the Broad Institute, aims to catalogue gene essentiality across hundreds of cell lines. Each Achilles gene set corresponds to a single cell line and contains the genes whose genetic knockout impacts the overall fitness of that cell line [31]. From Achilles, 75 sets containing at least

70 genes were selected at random for use with GeneSet MAPR. As with MSigDB, UP and DN segments were combined into a single gene set.

5.2.2 Ten Supporting Collections

As KnowEnG grew, gene sets from many other sources were made available to the project. Ten supporting collections of gene sets were selected semi-randomly from these growing databases, with an effort to avoid data that was already explicitly a part of the network used in this paper. For each of the ten chosen data sources, 40 gene sets within the size range of 100-2,000 genes were selected at random. This size restriction is meant to represent the typical scale of a user's input gene set while ensuring a valid, comparable AUC value.

For these collections, any sets represented by separate groups of up- and down-regulated genes are treated as two separate sets. That is, the UP and DN groups are not combined into a single set as is the case for MSigDB and Achilles. In fact, this separation is done explicitly for the LINCS collection, where only the down-regulated portions of gene sets are considered.

Three of these sets appear to be outliers. Both the Enricher Pathways and Reactome collections scored exceptionally high AUCs, suggesting some overlap with the data already in the network. For these two collections, the pathway subnetwork from Kegg was the third- and fourth-best performing network, respectively, as seen in Figure 5.3. However, both collections showed improved performance across most subnetworks, as compared against the remaining collections. Meanwhile, the other outlier is the collection of sets derived from Allen Brain Atlas, whose connections appear to be under-represented within the network. The Allen Brain collection is somewhat unique from the rest in that it attempts to map gene expression across the human brain.

5.2.3 One Sanity Check

Finally, one collection of gene sets was created with the explicit intention to overlap with data already in the network. For this test collection, 40 gene sets were created from the same GO terms that were used to create the GO

Table 5.2: Comparison of Methods by AUC Value

Collection	LASSO ensemble	DRaWR	GeneSet MAPR
MSigDB	0.695	0.684	0.758
dbGaP	0.593	0.622	0.677
Achilles	0.645	0.612	0.714
Allen Brain	0.528	0.578	0.644
Enrichr Path	0.951	0.960	0.966
Enrichr Pheno	0.737	0.774	0.806
ESCAPE	0.637	0.675	0.739
GeneSigDB	0.717	0.691	0.769
GEO	0.666	0.688	0.744
LINCS DN	0.705	0.642	0.745
Pathcom	0.684	0.679	0.756
Reactome	0.933	0.984	0.960
TargetScan	0.702	0.664	0.753
GO (test)	0.971	0.945	0.952
mean	0.726	0.729	0.785

subnetworks. These sets were not meant to gauge the overall performance of GeneSet MAPR; they were expected to perform well under the set membership prediction evaluation. Indeed, in Figure 5.3 it can be seen that they largely performed as expected. They also showed an unusually high utility for the genetic interaction subnetwork, an interesting aside which is less relevant to this paper than to those assembling the networks.

5.3 Comparison to Non-Network Approach

To tease out the effect of network structure on set membership prediction, we first compare GeneSet MAPR to an approach we will call *LASSO ensemble*. The LASSO ensemble approach applies the same standardization, sampling, and multiple regression voting as MAPR, but does not compute meta-path connectedness and otherwise ignores the network structure. Instead, LASSO ensemble constructs feature vectors where each individual node in the network — gene or annotation term — is treated as an unabstracted feature. If a gene is annotated by a given term, the value in that entry is set to the weight of the connecting edge, otherwise it remains zero. If a gene shares one or more edges with another gene, the sum of those edges is placed into that entry. Thus, each gene is represented by a sparse vector of over 40,000

individual features. In order to treat each subnetwork with equal emphasis, the weight of each edge in that subnetwork is divided by the sum of all weights, such that the total sum of all edge weights in a subnetwork is equal to 1. Thus, if one subnetwork contains many more edges than another, the weight of each individual edge is less for the denser subnetwork. The non-negative values of this feature vector are calculated from the same subnetworks used by the full MAPR method, and indicate the strength of the relationship between a gene and any other node on a one-to-one basis. In terms of the network, this approach considers every individual edge, rather than the pattern of occurrence of multi-length paths.

Despite utilizing the same raw data, the LASSO ensemble approach significantly underperformed the full GeneSet MAPR. Looking at the mean AUC over each gene set collection in Table 5.2, MAPR showed an average improvement over LASSO ensemble of 9.2%. Considering only the three benchmark collections — MSigDB, dbGaP, and Achilles — LASSO ensemble ranked 45%, 2%, and 15% of the sets in each respective collection with an AUC at or above 0.7. Meanwhile, MAPR ranked 85%, 22%, and 69% at 0.7 or better, respectively. These results suggest that the abstracted meta-path features, which consider network structure, are better at recovering gene set membership than specific observations on individual genes. It is likely that the meta-path features reduce the likelihood of over-fitting a model to noise in the data. Although the generation of meta-path features in MAPR requires more upfront computation, performing set membership prediction on the reduced number of high-level features makes the models more efficient and robust.

5.4 Comparison to DRaWR

Next, GeneSet MAPR's gene similarity ranking is compared to that which is output by Discriminative Random Walk with Restart (DRaWR), a state-of-the-art approach that has been shown to be effective in this context [41]. Like MAPR, DRaWR explicitly uses the network structure to define similarity of individual genes to an input set. Unlike MAPR, it does so without distinguishing between edge types, combining any heterogeneous subnetworks into a single homogeneous network.

5.4.1 Discriminative Random Walk with Restart

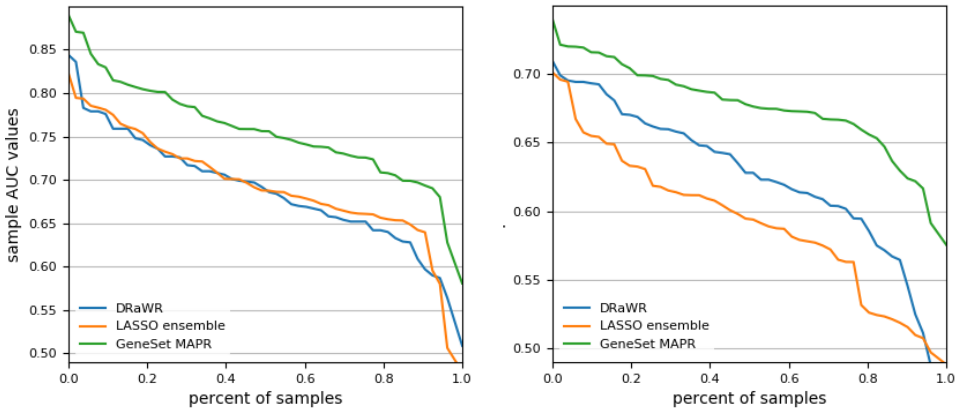
In a network, a *random walk* is a walk wherein at each node, the next edge in the walk is chosen at random from any of the edges shared by that node. A similarity measure can then be defined as the probability, if a random walk was initiated at a given source node, that the walk would terminate at any other specified node. A restart probability can be added to the metric, indicating the rate at which the walk will randomly be terminated and restarted from the source node. A higher restart rate, then, emphasizes shorter walks.

DRaWR applies the random walk approach, using the full input gene set as source nodes. The entire network is formed into a matrix, including non-gene terms, and the restart rate is an adjustable parameter. When subnetworks are combined, they are equalized by dividing each edge weight by the total sum of edge weights for the subnetwork. Weights are then summed across subnetworks, and compiled into one homogeneous network. The network matrix is multiplied by a weight vector until convergence. This is a key difference with MAPR, which specifically distinguishes the different meta-path types, quantifies connectedness for each separately, and then learns and uses the most relevant meta-paths to calculate the final ranking.

DRaWR was run by combining the eleven subnetworks from KnowEnG and setting the restart probability to 0.3. DRaWR was also tested with a restart set to 0.5, but this did not perform as well and is not presented here.

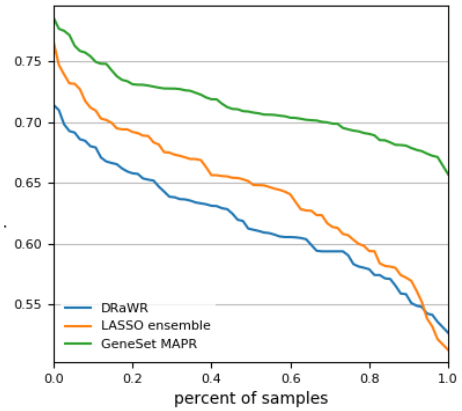
5.4.2 Results

MAPR achieved a higher average AUC per collection than DRaWR across all tested collections except Reactome, as seen in Table 5.2. In Figure 5.1, it can be seen that GeneSet MAPR performs well (AUC ≥ 0.7) on a higher percentage of sets from each collection than DRaWR. Specifically, MAPR performs well on 22, 10, and 50 more sets than DRaWR in MSigDB, dbGaP, and Achilles, respectively. The scatterplots of Figure 5.2 show the AUCs of individual sets, with the results from MAPR along the y-axis and DRaWR along the x-axis, where a dot along the main diagonal $y = x$ indicates the set acheived the same AUC using both methods. As shown in the figure, MAPR improved the AUC over DRaWR for 96%, 88%, and 100% of sets in MSigDB, dbGaP, and Achilles.



(a) MSigDB

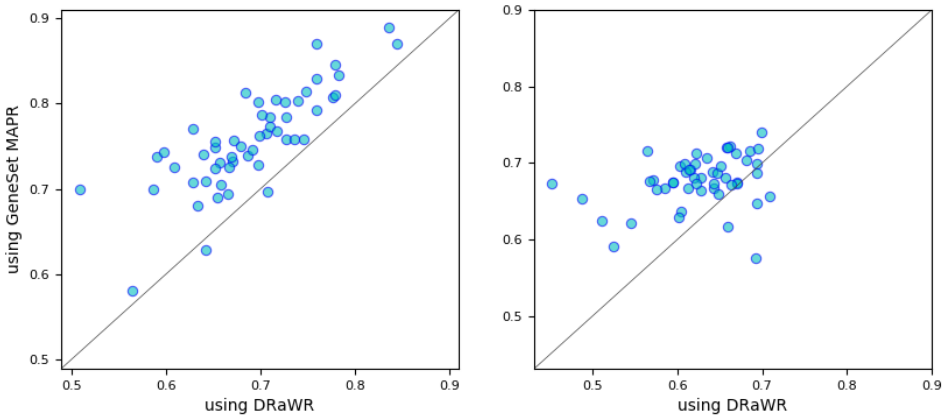
(b) dbGaP



(c) Achilles

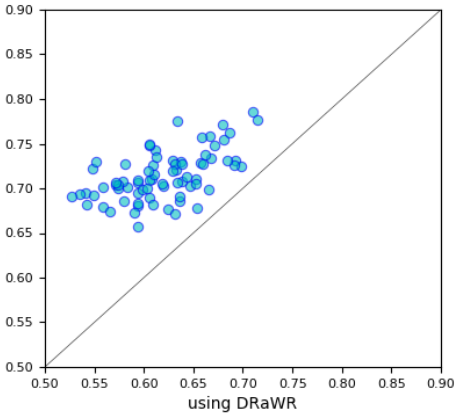
Figure 5.1: Comparison showing proportion of AUC scores for all sets across a collection. Plot shows the percentage of sets along the x-axis that achieved the AUC along the y-axis or better.

Although both methods rely on the same underlying subnetworks, the overall improvement of MAPR over DRaWR is likely due in large part to the consideration of the heterogeneous nature of the network through the use of meta-paths. Comparing all three methods (Figure 5.1) we see no improvement across MSigDB for considering the network structure in DRaWR versus LASSO ensemble, but notable improvement with the addition of the meta-path based features of MAPR. This is different from the dbGaP collection, where improvement can be attributed to the network structure (DRaWR vs. LASSO ensemble), with an additional boost due to the meta-path approach of MAPR. Additionally, as is apparent in Figure 5.2, there is some positive correlation between the DRaWR and MAPR AUCs, with a Pearson correlation of 0.80, 0.32, and 0.58 for MSigDB, dbGaP, and Achilles. This



(a) MSigDB

(b) dbGaP



(c) Achilles

Figure 5.2: Scatterplot of AUC scores for individual sets in each benchmark collection. X-axis represents the AUC received using DRaWR and y-axis the AUC from GeneSet MAPR. Sets along the diagonal received the same AUC for both methods, while those above performed better when using MAPR.

suggests that the connectedness of some gene sets is inherently higher than others when using the set of subnetworks on which DRaWR and MAPR were tested.

5.5 Effect of Individual Subnetworks

One idea we wished to explore was the effect of individual subnetworks on the performance of GeneSet MAPR. For each gene set, MAPR was run using only meta-paths created from a single homogeneous subnetwork at a time. Meta-

	MSigDB	dbGaP	Achilles	Allen Brain	Enrichr Path	Enrichr Pheno	ESCAPE	GeneSigDB	GEO	LINCS DN	Pathcom	Reactome	TargetScan	GO (test)	number of edges
full network	0.76	0.68	0.71	0.64	0.97	0.81	0.74	0.77	0.74	0.75	0.76	0.96	0.75	0.95	147,557,662
GO Bio Proc	0.69	0.64	0.66	0.62	0.91	0.74	0.69	0.69	0.69	0.68	0.72	0.90	0.69	0.88	16,773,233
Textmining	0.71	0.64	0.60	0.59	0.89	0.75	0.69	0.69	0.68	0.66	0.71	0.86	0.68	0.81	354,931
GO Cel Comp	0.67	0.61	0.66	0.62	0.81	0.68	0.66	0.68	0.68	0.66	0.67	0.77	0.67	0.83	73,632,134
GO Mol Func	0.66	0.61	0.65	0.61	0.84	0.68	0.66	0.66	0.66	0.65	0.68	0.77	0.66	0.77	48,104,940
PPI physical	0.65	0.60	0.64	0.57	0.80	0.69	0.65	0.67	0.67	0.67	0.67	0.80	0.70	0.72	184,435
PPI direct	0.65	0.58	0.64	0.56	0.80	0.66	0.62	0.67	0.65	0.67	0.66	0.76	0.67	0.76	86,569
Pathway	0.62	0.54	0.62	0.54	0.86	0.68	0.59	0.64	0.62	0.63	0.60	0.79	0.58	0.72	2,292,622
Protein Fam	0.61	0.61	0.58	0.55	0.80	0.62	0.62	0.59	0.57	0.56	0.62	0.77	0.65	0.52	5,423,388
Homology	0.56	0.57	0.54	0.53	0.80	0.59	0.58	0.55	0.54	0.53	0.59	0.83	0.58	0.75	652,823
Coexpression	0.52	0.51	0.51	0.50	0.51	0.51	0.52	0.51	0.54	0.52	0.50	0.58	0.50	0.54	50,100
PPI genetic	0.50	0.51	0.51	0.50	0.56	0.51	0.50	0.50	0.51	0.50	0.51	0.55	0.51	0.80	2,487

Figure 5.3: Comparison of mean AUC values over a collection for each subnetwork. The color range from red to green indicates the range of mean AUC values relative to that specific collection, from low to high. The total count of all edges in the network appear in the right-most column. Subnetworks are sorted by mean AUC over all collections.

paths of length one, two, and three were all still considered, but the edge type was the same at each step in the path. In Figure 5.3, the subnetworks are sorted by their average AUC across each collection. As can be seen, the full heterogeneous network outperforms any single homogeneous subnetwork across every collection. That is, MAPR successfully combines useful data from the assembled subnetworks such that the whole works better than its individual parts. Figure 5.4 shows a more visual representation of the relative results over the three benchmark collections, where it can be seen that some subnetworks had a much wider range of values than others for individual gene sets, and some subnetworks consistently outperformed others.

A few other trends are observable. The biological processes subnetwork is almost universally the best-performing subnetwork for each collection, while coexpression and genetic interaction are almost universally the worst-performing. As can be seen, however, coexpression and genetic interaction are incredibly sparse in this version of KnowEnG. In a network of over 23 thousand genes, there are over 500 million possible gene-gene pairs, and a subnetwork of a few thousand edges has very few opportunities to provide a meaningful signal to the model. On the other hand, the direct interaction

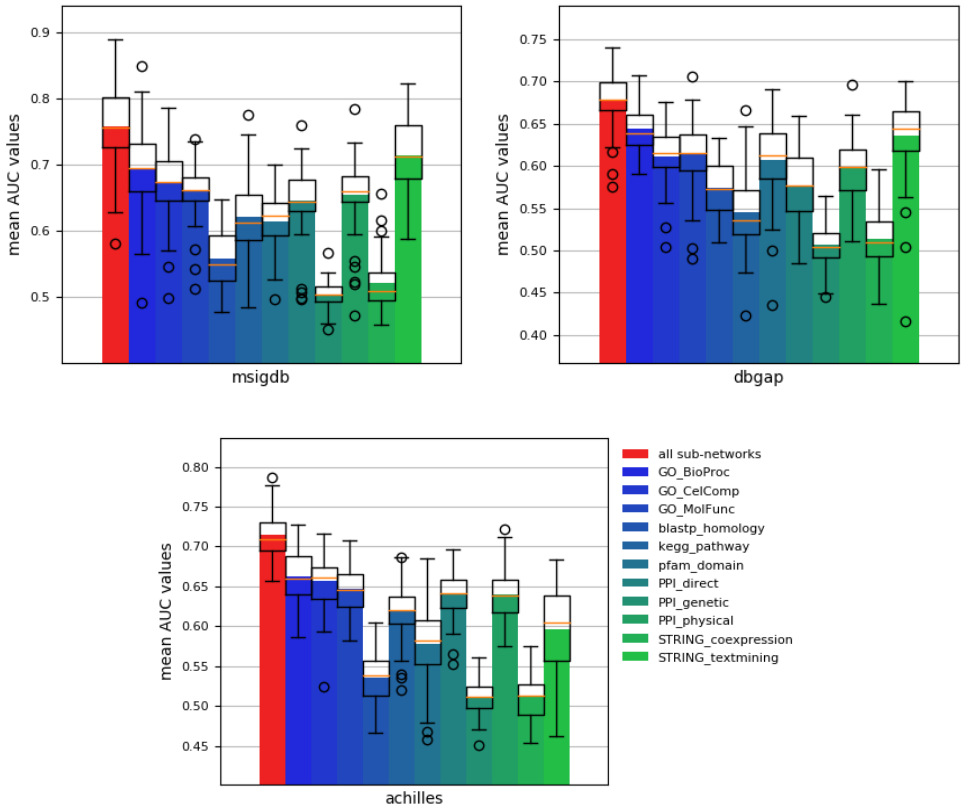


Figure 5.4: Bar graph shows mean AUC over a collection for subnetworks indicated in legend. The overlaid box plot shows the range for the upper and lower quartiles of all considered sets, with a horizontal line within showing the median. The whiskers extend from the median to include the farthest datum within a range of 1.5 times that of the the quartile.

subnetwork performs admirably for how sparse it is, with a typical AUC of 0.67 across each collection. It is reasonable to expect a dense network to perform better than a sparse one; however, the relationship between the typical collection AUC and the number of edges in a subnetwork is only moderately positive, with a Pearson correlation coefficient of 0.39. Here, MAPR's approach of standardizing meta-path counts relative to the expectation is able to compensate for varying edge density, putting more emphasis on the content and methodology of the subnetworks.

One surprising result was the performance of the textmining subnetwork. Textmining was often one of the best subnetworks across a collection, despite its relatively small number of edges and unorthodox premise. A clue may lie in how well it performed on the final test collection, the GO annotations.

Naturally, the GO collection received high AUCs with the GO subnetworks, but here textmining also performed quite well, suggesting there may be some overlap in the information conveyed by the GO subnetworks and textmining.

5.6 Meta-Path Rankings across Gene Set Collections

Similar to the effect of individual subnetworks, we can examine the utility of specific meta-paths within the network. Here, we examine the most useful meta-paths across the three main curated gene set collections. Specifically, this evaluation considers what percentage of sets within a collection used a given meta-path as one of the top five most uniquely identifying meta-paths. That is, we count how often a meta-path was given one of the five highest weights by GeneSet MAPR's feature importance list.

Tables 5.3, 5.4, and 5.5 show the ten meta-paths occurring most frequently in the top five for the MSigDB, dbGaP, and Achilles collections, respectively, along with the mean feature weight across the collection. For MSigDB, the majority of sets highly weighted a homogeneous textmining meta-path, and nearly half a homogeneous biological processes GO meta-path. This is unsurprising, as these two subnetworks were the best-performing subnetworks for MSigDB, as seen in Section 5.5. However, in contrast to the other two collections, sets across MSigDB are dominated by a few highly weighted features as seen by the high average feature weight and top-5 occurrence in Table 5.3. The top meta-paths for dbGaP and Achilles have lower rates of occurrence, indicating a more even distribution of selected meta-paths across the collections. In general, cellular component, biological processes, and textmining were very useful subnetworks and occur in various combinations with other edge types in the meta-paths for all three collections. However, whereas sets in dbGaP tend to use textmining in combination with other edge types, Achilles favored combinations utilizing GO, including an edge type omitted by the other two: molecular function. So while there is some similarity in the feature profile over a collection — with MSigDB appearing particularly homogeneous — there is variety in the meta-paths favored by each.

One further observation is the variation in meta-path length. While the thirty features presented in these three tables include fourteen length-3 meta-paths, the most frequently used paths are the shortest. Longer paths appear

Table 5.3: Important Meta-Paths for MSigDB

Weight	in Top-5	Meta-Path	Length
0.385	58.5 %	STRING_textmining-STRING_textmining	2
0.311	45.3 %	GO_BioProc	1
0.236	37.7 %	GO_CelComp	1
0.219	37.7 %	STRING_textmining-STRING_textmining-STRING_textmining	3
0.212	35.8 %	GO_BioProc-GO_CelComp	2
0.157	20.8 %	STRING_textmining	1
0.173	18.9 %	STRING_textmining-blastp_homology-STRING_textmining	3
0.148	15.1 %	PPI_physical_association-PPI_physical_association	2
0.098	13.2 %	pfam_domain-GO_BioProc-pfam_domain	3
0.116	11.3 %	pfam_domain-STRING_textmining-pfam_domain	3

Table 5.4: Important Meta-Paths for dbGaP

Weight	in Top-5	Meta-Path	Length
0.132	27.5 %	GO_BioProc	1
0.121	21.6 %	GO_CelComp	1
0.113	19.6 %	STRING_textmining	1
0.105	15.7 %	GO_BioProc-GO_BioProc	2
0.087	15.7 %	STRING_textmining-STRING_textmining	2
0.113	13.7 %	STRING_textmining-pfam_domain-STRING_textmining	3
0.095	13.7 %	STRING_textmining-STRING_textmining-STRING_textmining	3
0.092	13.7 %	blastp_homology-STRING_textmining-blastp_homology	3
0.064	13.7 %	STRING_textmining-GO_BioProc-STRING_textmining	3
0.082	11.8 %	blastp_homology-PPI_direct_interaction-blastp_homology	3

Table 5.5: Important Meta-Paths for Achilles

Weight	in Top-5	Meta-Path	Length
0.177	34.0 %	GO_BioProc	1
0.165	30.2 %	GO_BioProc-GO_CelComp	2
0.123	22.6 %	GO_CelComp	1
0.110	17.0 %	GO_BioProc-GO_BioProc-GO_CelComp	3
0.091	17.0 %	GO_BioProc-PPI_genetic_interaction-GO_BioProc	3
0.090	17.0 %	blastp_homology	1
0.068	11.3 %	GO_BioProc-GO_BioProc	2
0.062	11.3 %	GO_MolFunc-GO_CelComp-GO_MolFunc	3
0.061	11.3 %	GO_MolFunc-PPI_genetic_interaction-GO_MolFunc	3
0.061	11.3 %	GO_MolFunc-PPI_physical_association-PPI_physical_association	3

to play more of a supporting role in the models. At the same time, there is a greater number of longer meta-paths, with subtler variety. It may be that there is little difference in a model choosing to weight the meta-path m_i = (Ho, Bp, Ho) over m_j = (Ho, Tm, Ho) or m_k = (Bp, Ho, Ho). An evaluation focusing explicitly on path length appears in Section 5.7, and thoughts about selecting the most representative meta-path from a group appear in Chapter 7.

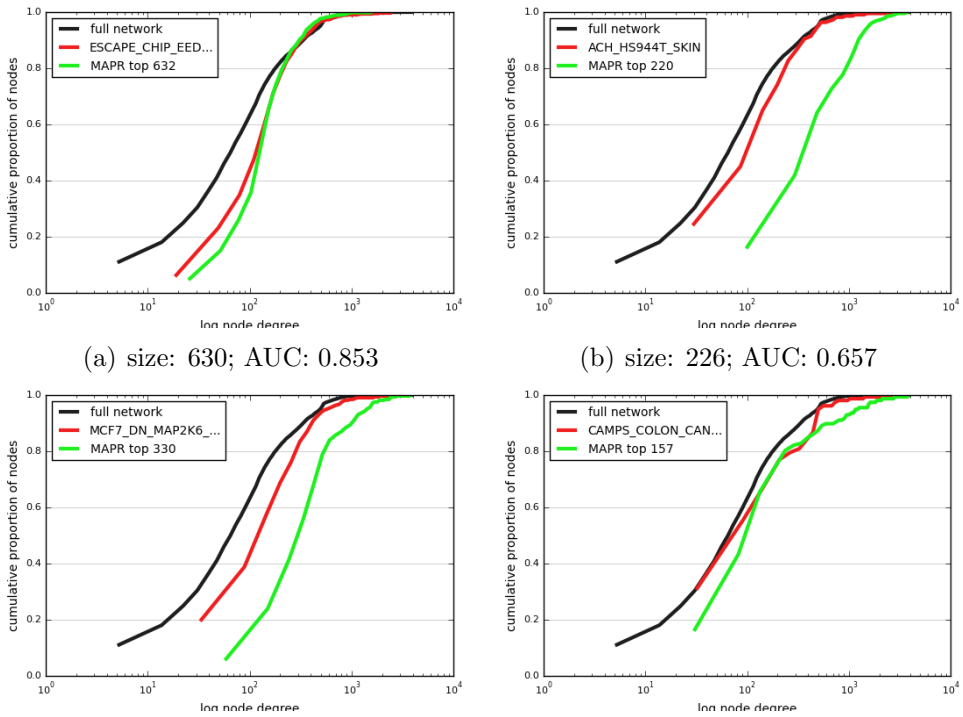
5.7 Consideration of Meta-Path Length

Table 5.6: Effect of Meta-Path Length on AUC Value

Collection	only Len=1	only Len=2	only Len=3	Len=1 or Len=2	all Lengths
MSigDB	0.731	0.749	0.751	0.752	0.758
dbGaP	0.671	0.672	0.661	0.674	0.677
Achilles	0.696	0.716	0.712	0.716	0.714
Allen Brain	0.643	0.642	0.640	0.646	0.644
Enrichr Path	0.959	0.964	0.955	0.967	0.966
Enrichr Pheno	0.794	0.799	0.788	0.806	0.806
ESCAPE	0.723	0.734	0.731	0.737	0.739
GeneSigDB	0.748	0.766	0.762	0.770	0.769
GEO	0.728	0.739	0.735	0.743	0.744
LINCS DN	0.715	0.747	0.741	0.748	0.745
Pathcom	0.734	0.752	0.751	0.753	0.756
Reactome	0.959	0.950	0.947	0.959	0.960
TargetScan	0.723	0.746	0.746	0.748	0.753
GO (test)	0.950	0.935	0.923	0.951	0.954
mean	0.770	0.779	0.775	0.784	0.785

Another aspect of GeneSet MAPR we wished to understand was the effect of path length. The longer a path, the less intuitive sense it is likely to offer. So to better understand the effect of path length, MAPR was run over all gene sets with varying restrictions on path length. The approach used in the above evaluations considered meta-paths with a length of up to 3 edges. MAPR was then run where it was restricted to using meta-paths of length 1. Two additional runs were performed, using length 2 and length 3, respectively. Finally, a run was performed allowing a mixture of lengths 1 and 2, but nothing longer.

As seen in Table 5.6, using meta-paths of only length 1 resulted in the



(a) size: 630; AUC: 0.853

(b) size: 226; AUC: 0.657

(c) size: 336; AUC: 0.849

(d) size: 165; AUC: 0.690

Figure 5.5: Plots show proportion of nodes with degree of X or less for an input set and the corresponding top N genes returned by GeneSet MAPR. The x-axis is presented in log scale to account for the scale-free nature of the network.

lowest average collection AUC. Meta-paths of length 2 performed better than length 1, and the combination of lengths 1 and 2 resulted in yet another boost in performance. However, the addition of length 3 meta-paths resulted in a flat or very modest improvement for most collections, and in some cases a slight decrease.

It appears that increasing the meta-path length has diminishing returns. As discussed earlier, separating the gene ontology subnetwork into three separate subnetworks improved performance. This suggests that, in striking a balance between variety of edge types and length of meta-paths, there may be more to gain from either adding additional subnetworks or logically separating large ones than merely adding longer meta-paths.

5.8 Distribution of Node Degree in Results

As mentioned in Chapter 3, the scale-free nature of biological networks poses a challenge when attempting to define similarity. GeneSet MAPR employs two steps to mitigate this issue, as outlined in Chapter 4. First, the transition matrix adjusts the likelihood of one gene connecting to another according to the originating gene node's overall edge count. Second, standardization of meta-path counts accounts for meta-paths with significantly higher or lower density. Here, we examine how well MAPR meets this challenge.

Figure 5.5 shows comparisons of four sets: two where the top-ranked genes output by GeneSet MAPR show a similar distribution of node degree as the original set, and two where MAPR's results are skewed towards higher-degree nodes. These sets were chosen semi-randomly, to represent large and small sets, with high and low AUCs. Each figure shows the cumulative distribution of nodes by degree for the full network, the original set, and the top-N ranked genes from MAPR, where N is the same size as the original set.

Data for all sets evaluated in this chapter can be found in Table A.1. For every set, the median degree of the original set is compared to the median degree of the top-N MAPR-ranked genes. In another column, the difference between the two medians is adjusted to show the change relative to the median of the original set. As another way of presenting the data, there is also a comparison of the proportion of genes with a degree greater than 200, and the change from the original set to MAPR's top-N. (As mentioned in Chapter 3, the median node degree in the network is 69.)

Most sets showed some increase from the median degree of gene nodes in the input set to the median degree in the MAPR top-N. A small proportion of sets showed a significant increase in median node degree. However, while the median degree of the input set was related to the AUC achieved by MAPR, with a Pearson correlation coefficient of 0.64, the amount by which the median degree changed showed a slightly negative correlation to MAPR AUC (-0.34). That is, MAPR tended to perform worse on sets where its results improperly skewed towards higher-degree nodes. Interestingly, set size shows little correlation with any of the other factors, and only a very weak negative correlation (-0.12) with set membership prediction AUC. Rather, the single item that correlates the most with set AUC is the density of edges in the set. This correlation is only moderate, but it does point to an unfortunate, if

somewhat obvious, conclusion. The sets for which GeneSet MAPR can best predict hidden members are those with the most knowledge upon which to draw, in the form of network connections. (This is true for DRaWR as well, which used the same KnowEnG network. The correlation between DRaWR AUC and median node degree for a set was 0.58.) Naturally, the better a profile that can be created for a set using the underlying network, the more reliable the results.

CHAPTER 6

CHARACTERIZATION OF A NOVEL GENE SET VIA GENESET MAPR

To illustrate the utility of GeneSet MAPR as a gene set characterization tool, here we report a detailed analysis of a recently developed gene set. The set comes from the Breast Cancer Genome Guided Therapy Study (BEAUTY) [42], and was created by examining the differential expression among patients diagnosed with a form of breast cancer that is particularly resistant to treatment. MAPR was used to rank genes by similarity to the set — to identify the interaction neighborhood — and to help characterize the make-up of the set. Of specific interest are genes ranked highly by MAPR, but which failed to cross the study's statistical thresholds, or otherwise went unmentioned. MAPR's ability to learn the pattern of connectedness within a set and identify related genes is especially useful here, where hard cutoff values established by convention are susceptible to noise over a small sample size.

Indeed, sample size was a challenge in the study as a whole, as it is for any study whose subjects suffer from a rapidly evolving disease. Additionally, the gene set of interest comes from a subset of the study consisting of only 44 patients. With this sample size and the desired p-value specified at 0.05 or less, the statistical power of this subset is greatly reduced, increasing the likelihood of a Type II error. A Type II error occurs when an effect should be witnessed but is not. In this case, a Type II error would result in the omission of a gene from the set when it should be included — the perfect opportunity for a tool such as GeneSet MAPR to fill in missing members of the set.

6.1 BEAUTY Triple Negative Responders: A New Gene Set

We were presented with a gene set curated as part of the longitudinal BEAUTY study performed at the Mayo Clinic [42]. The set consists of genes exhibiting significant differential expression (DE) levels between responders and non-responders diagnosed with triple negative breast cancer (TNBC), a clinical molecular subtype of breast cancer where tumors are negative for ER (estrogen receptor), HER2 (epidermal growth factor receptor 2), and PR (progesterone receptor). This gene set will be referred to as the BTNR set.

BEAUTY consisted of 132 patients with stage I-III breast cancer, with a tumor at least 1.5 cm in size, who had been recommended for treatment by neoadjuvant chemotherapy. Surgery was performed following completion of a targeted drug regimen. Patients experiencing pathological complete response (pCR) were identified as those who had no invasive tumor in the breast and axillary lymph nodes. A third of the patients had TNBC and experienced a pCR rate of 54.5% (24 out of 44), as opposed to the overall rate of 33.3% (44 out of 132). The log fold change (logFC) in DE of genes in tumor tissue was measured between pCR and non-pCR TNBC patients. Genes in the set met two thresholds: p-value 0.05; and the absolute value of logFC 1. The final BTNR set consisted of 384 genes, of which 323 occur in the KnowEnG network with at least one connecting edge.

MAPR emphasizes connections to genes which these statistical thresholds alone could not. One highly ranked group identified in BTNR are *claudins*, many of whose members are markers of poor survival in cancer patients, but who show only subtle changes in DE. Two other groups are *kallikreins* and *collagen alpha chains*. While a few of these genes are part of the BTNR set, they were not mentioned in the BEAUTY study; MAPRs similarity ranking brought attention to them. Ranking and set membership is contrasted against a handful of other cancer sets selected from the MSigDB collection. The rates of single nucleotide variation (SNV) and copy number variation (CNV) between pCR and non-pCR patients is provided in the BEAUTY supplemental tables.

Gene mutation rates from The Cancer Genome Atlas (TCGA) are also examined, drawing on over 30,000 cases. Using TCGAs Genomic Data Commons Data Portal (GDC), the mutation rate is calculated as the percentage

of all cases in GDC where the gene was tested for simple somatic mutations that were found to be positive. Additionally, the Catalogue of Somatic Mutations in Cancer (COSMIC) maintains a cancer gene census, cataloguing genes causally implicated in cancer. Genes in Tier 1 show documented relevant activity and evidence of mutations that promote oncogenic transformation. Tier 2 genes show strong implications in cancer but with less evidence, typically indicating more recent targets.

6.2 Set Characterization from MAPR Feature Ranking

Table 6.1: Meta-Path Importance for BTNR

Rank	Weight	Meta-Path	Length
1	0.497	GO_CelComp	1
2	0.362	GO_BioProc	1
3	0.333	STRING_textmining-blastp_homology-STRING_textmining	3
4	0.308	STRING_textmining-kegg_pathway-STRING_textmining	3
5	0.239	blastp_homology-GO_MolFunc-blastp_homology	3
6	0.228	PPI_physical_association-PPI_physical_association	2
7	0.162	GO_BioProc-GO_BioProc	2
8	0.161	PPI_genetic_interaction-blastp_homology	2
9	0.157	blastp_homology-STRING_textmining-blastp_homology	3
10	0.140	pfam_domain	1
11	0.133	blastp_homology-GO_BioProc-blastp_homology	3
12	0.124	pfam_domain-PPI_genetic_interaction-pfam_domain	3
13	0.120	PPI_physical_association-blastp_homology	2

The feature ranking output by GeneSet MAPR typically provides around 10-20 positively weighted meta-paths. There is typically a handful of meta-path features with dominant weights, and a long tail of lower-weighted paths which may be variations on a given pattern. The higher the number of regression models specified by the user, the longer the tail of minuscule aggregated feature weights. Here, we only consider features with a weight of 0.10 or greater.

Table 6.1 shows the top-ranked meta-paths for BTNR. As can be seen, the two most dominant meta-paths are the length-1 paths GO_CelComp and GO_BioProc. These two meta-paths consider connections through any of the GO terms in the GO hierarchy that appear under cellular component and biological processes, respectively. This indicates that the genes in BTNR

were most connected to each other through shared membership in these two collections of GO terms.

The next two most important meta-paths contain the STRING_Textmining edge, leveraging shared presence in academic literature. While a shared textmining edge offers less biological insight, textmining has been one of the more predictive edges for set membership over the tested gene set collections as it does indicate a body of evidence exists relating one gene to another. Along these lines, the high occurrence of blastp_homology in these top-ranked meta-paths indicates that our knowledge of direct functional relationships between genes in BTNR may be limited, and in this case connections are facilitated through homologous genes.

6.3 Novel Findings from MAPR Gene Ranking

GeneSet MAPR was applied to rank genes by connectedness to BTNR. Of specific interest are genes found by MAPR to have high connectedness to the set, but which failed to meet BTNR's statistical cutoff values for DE and p-value. The following three families of genes showed a dramatically increased representation among MAPR's top-ranked genes. Data regarding mutations and presence in other cancer sets can be found in the referenced appendices.

We focus the evaluation of GeneSet MAPR's ranked gene list on the top 400 genes. This number was chosen somewhat arbitrarily to represent the top-N genes a researcher may consider worth her time to investigate. In a network of 23,782 human genes, 400 represents merely the top 1.7%. Additionally, 400 is similar in scale to the original BTNR set (384 genes).

Claudins

Claudins are important to cell adhesion and flow of molecules in the intercellular space. While BTNR contains 6 claudins, MAPR lists 20 with a rank of 400 or better, as seen in Tables B.1 & B.2. Three of these were found by the BEAUTY study to have a higher rate of SNV or CNV mutation in TNBC non-responders, yet were not captured by the BTNR set. Claudins show a low rate of inclusion in the other 9 cancer sets and are not highly ranked by MAPR for those sets, suggesting a unique relevance to BTNR.

Claudins have been implicated in the progression of metastasis in various cancers, as they play a critical role in allowing or blocking cancer cells from crossing endothelial barriers [43]. Numerous epithelial-derived cancers display altered claudin expression patterns and certain claudins can now be used as biomarkers to predict patient prognosis. Changes in claudin expression patterns have been observed in numerous cancers, and some specific claudins can be used as biomarkers indicating patient prognosis. Indeed, claudin-low tumors have been identified as a distinct molecular subtype [44, 45]. This is facilitated by interactions with E-cadherin (CDH1), which was relatively highly ranked by MAPR (rank 494 of 23,782). Interestingly, while claudin-low tumors tend to display a triple-negative phenotype, very few triple-negative breast cancers are claudin-low.

Kallikreins

As serine proteases, kallikreins play a role in severing peptide bonds. In Tables B.3 & B.4, it can be seen that BTNR contains 7 kallikreins and MAPR lists those plus another 7 in the top 400. Most of BTNR's 7 kallikreins are members of the other explored cancer sets, while MAPR's additional kallikreins are not. One of these additions, KLK2, is listed as a Tier 1 cancer gene by COSMIC but was not included in BTNR by BEAUTY.

Kallikreins are cited by many studies as indicators of poor prognosis in cancer [46], and are often down-regulated in breast cancer patients. Indeed, many of the kallikrein family were found to have lower expression in BEAUTY non-responders than in responders. Several kallikreins have been shown to be useful biomarkers for breast cancer, including KLK3, KLK4, KLK5, KLK6, KLK8, KLK10, KLK13, and KLK14 [47]. Most of these were included in BTNR and other cancer sets, and MAPR lists all but KLK3 in the top 400. However, their presence was not addressed by BEAUTY [42]; MAPR's gene ranking drew attention to their potential importance.

Collagen Type N Alpha Chains

The BTNR set contains 6 collagens, all of which receive a MAPR ranking of at least 233, as shown in Tables B.5 & B.6. MAPR ranked an additional 31 in the top 400, most of which have potentially significant levels of mutation

among cancer patients, according to TCGA. Of these, 3 were observed by BEAUTY to have a significant difference in rate of SNV and/or CNV mutations between TNBC responders and non-responders, but were not included in BTNR. COL3A1, at rank 378, is listed as a Tier 2 gene by COSMIC, but was also not included in the BTNR set. COL1A1, at rank 411, is similarly listed as a Tier 1 gene, but was not included.

One gene each from the Type II, IV, VI, XXVIII collagen chains, as well as two from Type IX, were found to be differentially expressed between responders and non-responders. MAPR ranks many of the alpha-chains highly, with the Type IV, V, VI, XV, XVII, XXV, XXVII, and XXVIII chains falling within the top 100, and several others falling within the top 400. Several alpha-chains have been linked to cancer progression. Type VI (5 genes at rank 278 or better) promotes both tumor progression and chemotherapy resistance, suggesting a feasible anticancer strategy wherein collagen VI is suppressed [48]. Increased Type V (3 genes at rank 348 or better) has been found in the desmoplasia surrounding tumors in breast tissue [49], and evidence in mice indicates Type V chains as potential targets for inhibiting tumor growth [50]. Type X (rank 88) has been identified as a biomarker for colon cancer, with increased levels indicating the presence of a tumor [51].

Aberrations in Type I and Type III chains have been linked to malignant tumors via the formation of readily degradable collagen bundles [52]. Both Type I genes were highly ranked, with COL1A2 (rank 267) showing a notable difference in mutation rates between pCR and non-pCR TNBC patients. The Type III collagen COL3A1 (rank 378) showed a significant general rate of mutation and is listed as a Tier 2 gene by COSMIC. Types XIV (rank 155) and XXII (rank 59) showed significant rates of mutation among pCR TNBC patients, as opposed to non-pCR, although neither exhibited the required DE and p-value to be included in BTNR. Interestingly, Type XI alpha 1 (COL11A1) was ranked highly by MAPR (rank 170), while the remaining members of type XI were not. COL11A1 has been identified as an accurate marker for invasive breast carcinoma lesions [53], and as a potential target for general cancer therapy [54].

6.4 Enrichment Using MAPR Gene Ranking

One common approach to characterizing a gene set is to examine enrichment for annotation terms. Put simply, if an annotation term has more members in the target gene set, then the set is positively enriched for that term. Conversely, if the set contains fewer members of the annotation term, then the term is negatively enriched. Enrichment is not a purely binary concept; terms can be more or less enriched than others for a set. A popular tool for comparing enrichment is Gene Set Enrichment Analysis (GSEA), by the Broad Institute. GSEA compares random permutations of the input phenotype or gene set to return a normalized enrichment score (NES) for each annotation term considered.

In the case where a user has a ranked or otherwise scored list of all genes, the NES indicates how strongly members of a term are clustered near the top or bottom, or whether they are evenly distributed. This is compared against the null hypothesis — what would be expected in a uniform random distribution — by creating random permutations of the input set.

A ranked list for GSEA was created from BTNR by sorting the 14,228 genes for which BEAUTY had DE values by the magnitude of the fold change. There occurred 619 ties, which were ordered according to their p-value. The remaining 9,554 genes for which BEAUTY had no values but occurred in KnowEnG were tied at the bottom of the list. The MAPR ranked list contained 1,417 ties distributed fairly evenly throughout the 23,782 genes.

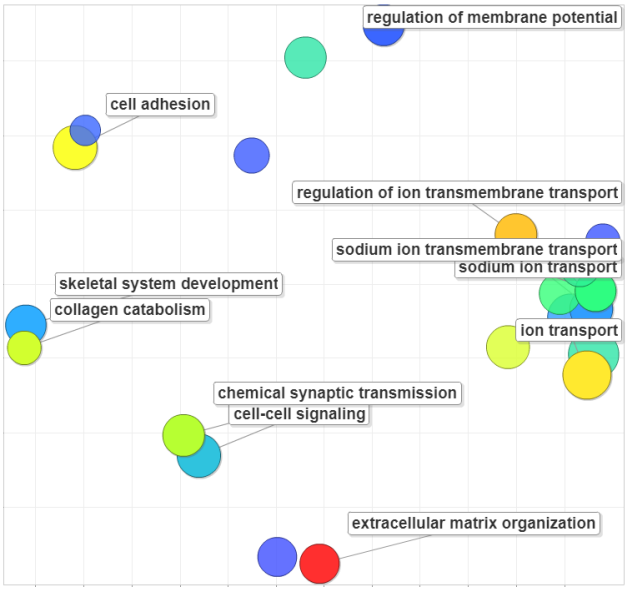
6.4.1 Results from GSEA

Using the MAPR ranked list as input, terms were sorted by the NES returned by GSEA, which can be found in Table C.1. Terms consisting of over 500 genes were ignored. Among the top terms, the following functions were highly represented: membrane / extracellular matrix, ion transport, cell adhesion, and signaling. The top 50 GO terms were selected and entered into ReviGO to create the visualization in Figure 6.1(a). ReviGO transforms the data into a lower-dimensional semantic space, enabling some of these functional clusters to be highlighted visually [55]. For example, a large cluster of transport terms can be seen to the right, and smaller clusters, such as the transmission/signalling pair, are scattered elsewhere. ReviGO's own

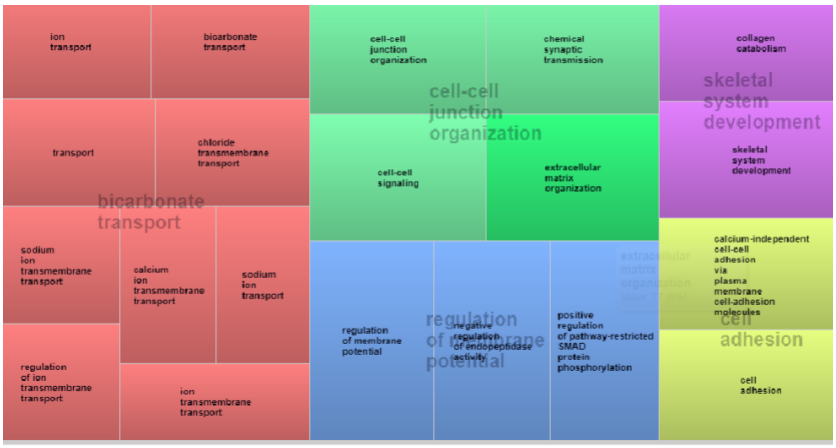
functional clustering of the GO terms can be seen in Figure 6.1(b).

Of particular interest are terms ranked highly by GSEA after running GeneSet MAPR that were not otherwise ranked highly for BTNR or the other 9 cancer sets — that is, uniquely important terms that were discovered only after running MAPR. These terms and respective GSEA scores appear in Table C.1. Of these, terms containing members of the gene families in Section 6.3 were related to protein digestion/absorption (via kallikreins' effect on peptide bonds), the extracellular matrix, or membranes. Top-scoring terms that largely did not contain members of those gene families typically related to signaling, epidermal growth factor, cell adhesion, or ion transportation across membranes.

One final question is how much the enrichment scores are influenced by the families of genes identified in Section 6.3. Put another way, one might have the reasonable concern that MAPR overfit its model to the input gene set, identifying a small handful of annotations that are highly enriched for the set and simply returning genes sharing those annotations. However, the approach of using meta-paths abstracts away some of these finer details, reducing the potential to overfit to noisy data. Looking at the top 100 terms returned by GSEA as positively enriched for the MAPR ranked list in Table C.1, only 38% contained any genes from the families discussed in Section 6.3. In fact, no single term accounted for more than 17% of the top 323 MAPR-ranked genes. So while MAPR's gene similarity ranking is influenced by the terms for which a set is enriched, such terms are clearly not the only factor, as might be expected if the model overfit to the terms in the network.



(a)



(b)

Figure 6.1: Top shows top-scored GO terms in ReviGO's reduced-dimensionality semantic space, where a few functional clusters can clearly be seen. Bottom shows ReviGO's interpretation, represented by the largest term in the group.

CHAPTER 7

FUTURE WORK

As with any research project, it seems the single most constricting resource is time. Several topics are outlined here where either an addition was planned, or an implementation had not undergone proper evaluation. Some interesting ideas were left aside in order to properly evaluate the core functionality of GeneSet MAPR. It is the hope of the author that we will have the opportunity to explore these ideas in the future.

One partially completed evaluation against another method is presented first. Next, the following potential additions are presented in order of expected effort to implement, from least to most.

7.1 Comparison to GeneMANIA

It was our intention to provide a comparison to another established state-of-the-art network-based method for ranking genes by similarity to a set: GeneMANIA [56]. However, as provided, GeneMANIA utilizes a much different set of underlying networks than was used to evaluate GeneSet MAPR, instead leveraging over a hundred relatively sparse networks. It is possible with GeneMANIA's source code to import our own subnetworks, allowing for a more direct one-to-one comparison of the methods. Unfortunately, this is a more drawn-out process that is harder to automate. While initial comparisons over a small handful of gene sets was positive, we felt it would be inappropriate to provide results until we could show trends over a broader sample size. It should be noted that in [41], DRaWR compared favorably against GeneMANIA, so we are optimistic.

7.2 Connectedness as Enrichment

It is reasonable to suppose that GeneSet MAPR's connectedness could be used as a stand-in for enrichment. MAPR already finds genes that share more connections to the input set than would be expected for a random sample of the same size; it is a trivial manner to apply the same process to individual annotation terms. Indeed MAPR already outputs a scored list of enriched terms. To do this, a gene-by-term feature matrix is created, to which the sampling and regression approach from Section 4.4 is applied. Any term sharing an edge with a gene in the set is considered, while any term not connected to the set cannot be scored, and anyway would never be considered enriched.

The connectedness score from MAPR was compared against normalized enrichment score from GSEA for several gene sets selected from those with the highest and lowest AUCs. The exact same annotation terms from the KnowEnG network were input into GSEA to ensure both methods used the same data. The correlation between MAPR and GSEA enrichment scores was surprisingly low. Curiously, GSEA managed to produce positive enrichment scores for some annotation terms that shared no edges with the input gene set. Furthermore, the enrichment scores from GSEA over each gene set were moderately correlated to the size of the annotation terms (typically $\sim$0.45). This correlation was even stronger when considering only positive enrichment scores (typically $\sim$0.76). The connectedness score from MAPR showed almost no correlation to term size (absolute value less than 0.23). Here, only immediate length-1 connections were compared, but this process can readily be generalized to longer paths, essentially unrolling a meta-path to see what terms facilitate the connections. Connectedness is a logical way to think about enrichment, and this aspect deserves further investigation.

7.3 Accepting Ranked Sets

Currently, GeneSet MAPR accepts a gene set as a simple list of items without any ranking or scores. Naturally, a researcher may already have a probabilistic understanding of their gene set, or some other measure by which the set has been ranked. Implementation of this change would be easy enough. In

fact, two options exist: weight the contribution of individual genes when calculating connectedness, or weight their importance when training the LASSO models. Evaluation of this change is more complicated. Potentially, calculation of the AUC score could be modified to accommodate a weighted input set. For example, the contribution of each hidden gene along the y-axis (true positive rate) could be changed from a 1-to-1 contribution to proportional dependent on the sum of weights of genes in the hidden fold. It is also quite possible that a more appropriate metric already exists. In either case, the concept must be evaluated before inclusion in the final product.

7.4 Improving the Classifier Model

Currently, GeneSet MAPR uses an ensemble of LASSO regression models as the classifier. (Technically, the term *classifier* is a little misleading, as MAPR is more interested in the score output by the classifier than the hard class assignment.) Other regression-based models were evaluated, such as logistic regression, support vector machines (SVM), and elastic net. However, none performed as well, and the drawback of more complicated models is that the importance of individual features — and with that, the intuition of a set is intrer-related — is often lost.

One option that would be good to explore is the use of independent component analysis (ICA) as an approach to feature reduction. Like LASSO, ICA identifies correlated features and can reduce the feature space to a small number of orthogonal features, which are weighted combinations of the original features, or mixing weights [57]. Unlike LASSO, the results can be deterministic. That is, from one run to another for two highly correlated features, LASSO may not zero-weight the same feature each time. The downside of ICA is that extracting individual feature importance from a combination of reduced-dimensionality mixing weights is a little more complicated, especially since ICA is likely to assign negative weights to many features. An alternative to ICA is non-negative matrix factorization (NMF), which produces purely additive mixing weights but requires the original feature values be non-negative [58]. Once these issues are addressed, either ICA or NMF could be swapped in, in place of the LASSO model.

Additionally, the framework was created to allow *bootstrapping*, where mul-

tiple classifiers are run in sequence. At each round of training, the classifier model sets aside those items which it can easily classify and passes the more difficult items on to the next classifier. One issue to be addressed here is how to avoid overfitting. Another is how to return feature importance, as some items will pass through multiple classifiers as others will not. Although the framework is in place, this approach has yet to be properly evaluated.

Finally, the idea was raised that if the genes within the training set could be clustered, then separate classifiers could be trained on those clusters. However, how to cluster a set of items in an unsupervised manner when it is unknown whether an appropriate clustering even exists is a body of research unto itself. We deemed that the random sub-sampling of training data in Chapter 4 adequately addresses this problem. Time permitting, this would be an interesting avenue to explore.

7.5 Clustering Meta-Paths

As mentioned in Chapter 5, there may be little intuitive difference between meta-paths m_i = (Ho, Bp, Ho), m_j = (Ho, Tm, Ho), or m_k = (Bp, Ho, Ho). The user may appreciate simply being shown that the group of meta-paths $M = m_i, m_j, m_k$ was uniquely important, as opposed to seeing the weights of the individual paths relative to each other. As such, an alternative to changing the classifier, such as implementing ICA or NMF, would be to group highly correlated meta-paths. Then, representative meta-paths from each group could be passed to the LASSO model, or else aggregated feature importance weights could be grouped under the corresponding representative meta-path. For example, once the meta-path features have been calculated for a gene set, hierarchical clustering could be used to define a few high-level "classes" of meta-paths. Then the feature importance score would highlight which classes of meta-path are most prominent within the set. Such an approach may be moot if the following section is implemented.

7.6 Selective Computation of Meta-Paths

It may be possible to identify meta-paths that are likely to be correlated before computing them. Note that this is slightly different than the previous section, which proposes grouping meta-paths by their correlation of importance to a gene set. Instead, this section proposes examining the components of a potential meta-path — the adjacency matrices — to compare how highly correlated they are across the entire network. If two edge types are highly correlated across the network, then the meta-path crafted from those two edge types is unlikely new information. Similarly, if a length-2 meta-path is roughly the same as an already-existing edge type, then we may wish to skip computing the resulting length-3 meta-path.

Indeed, we see in the feature importance lists of Chapter 5 that similar length-3 meta-paths with minor variations tend to be assigned similar weights. If we can decide upfront how to be selective in computing meta-paths, essentially pruning them, then we can further reduce the time required for the network pre-processing and feature generation steps.

7.7 Including Other Species

As originally envisioned, GeneSet MAPR would consider connections to genes outside the target species. Often, experiments can be performed on another species, such as mice, that cannot be performed on humans. What we might expect to see in this case is a human gene relate to a mouse gene through homology or orthology, that mouse gene relate to another through experimental evidence, and finally that second mouse gene relate back to a human gene. The result would follow a meta-path along the lines of $m = Ho, Ex, Ho$, where Ho is homology and Ex the edge type relating the experimental evidence.

A few options exist, each with trade-offs. The most straightforward would be to expand the gene-gene matrices to include mouse genes, increasing the burden on memory and computation. Another option would be to treat the non-human genes the same way as annotation terms, where the nodes are removed and edges are redrawn between pairs of human genes. Unfortunately, this removes the ability to connect directly from one non-human gene

to another. Finally, a compromise could be crafted where there exist intra-species matrices and inter-species matrices. When calculating meta-paths, then, the algorithm would need to be aware of when a meta-path crosses from one species to another. Such a consideration is feasible, but requires some changes to the network pre-processing code to allow for it.

7.7.1 Beyond Star Networks

A star network is one wherein there exists a central node type. Other node types connect to the central node, but never to each other. The network used in this paper is a star network, where gene nodes are the central node type, and various annotation nodes connect to those. However, as mentioned in Chapter 3, the GO subnetwork has its own hierarchy where annotation terms are interrelated. Including this would break the star network assumption.

The same compromise suggested to incorporate additional species above could also address the issue of inter-connections between annotation terms. In this case, three types of matrices would exist: gene-gene, a transitional gene-term, and term-term. Care would be taken when crafting the meta-paths to ensure the path began and ended on a gene node, and to select the appropriate transition and term-term matrices. This may require slightly more attention from the provider of the network to ensure the file is formatted properly, but the flexibility to allow term-term and inter-species connections may be worth it.

CHAPTER 8

CONCLUSION

In this thesis, we present GeneSet MAPR, an algorithm for finding items in a network that are similar to an input set while offering an intuitive description of what makes the set unique. We simultaneously show the utility of connectedness over meta-paths to identify unique patterns while addressing some common concerns about big data and machine learning approaches. The method used by MAPR, outlined in Chapter 4, shows an improvement over other methods at the task of set membership prediction, as seen in Chapter 5. Moreover, it accomplishes this while being readily adaptable to other and more complex tasks, several of which are detailed in Chapter 7. It is our belief that this flexibility is due largely to the treatment of the underlying subnetworks and their structure.

We provide a detailed example of an application to the genomics domain in Chapter 6. There we show how MAPR can leverage a wide array of established domain knowledge to emphasize connections and trends within a researcher's input set, helping to guide her towards more efficient discovery. This is especially useful in the case where a gene set is created from a small sample of patients, and a desire to use p-value as a threshold can lead to errors of magnitude: both a high likelihood of omitting relevant items and an exaggerated effect size for those that do meet the threshold. We show that MAPR is able to add context to the (sometimes under-appreciated) uncertainty of statistical analysis and draw attention to items where further research appears promising.

We show that meta-paths can be a valuable way of exploring patterns within a network. Further, we show it is possible to automatically identify which meta-paths are most unique within a set, removing from the user the burdens of preselecting important meta-paths, defining expected number or size of clusters, or otherwise seeding the analysis in some fashion. Of course, in the true spirit of ensemble learning we believe the method behind MAPR is

best used in conjunction with additional approaches to quantifying network patterns, such as random walks or clustering.

Given the uncertain and incomplete nature of our knowledge in the genomics domain, it is in some ways remarkable that GeneSet MAPR — or any method — performs as well as it does. We believe the use of network structure to capture multi-step connections shows the most potential for exposing hidden dependencies between items in this and many other domains.

REFERENCES

[1] P. M. Sommers, "The Super Bowl theory: Fourth and long," *The College Mathematics Journal*, vol. 31, no. 3, pp. 189–192, 2000.

[2] T. Aittokallio and B. Schwikowski, "Graph-based methods for analysing networks in cell biology," *Briefings in Bioinformatics*, vol. 7, no. 3, pp. 243–255, 2006.

[3] F. Ball and P. Neal, "Network epidemic models with two levels of mixing," *Mathematical Biosciences*, vol. 212, no. 1, pp. 69–87, 2008.

[4] S. Brin and L. Page, "The anatomy of a large scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1/7, pp. 107–17, 1998.

[5] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[6] Z. Yang, R. Algesheimer, and C. J. Tessone, "A comparative analysis of community detection algorithms on artificial networks," *Scientific Reports*, vol. 6, no. August, 2016. [Online]. Available: http://dx.doi.org/10.1038/srep30750

[7] J. Leskovec, K. J. Lang, and M. W. Mahoney, "Empirical comparison of algorithms for network community detection," *Proceedings of the 19th International Conference on World Wide Web*, pp. 631–640, 2010.

[8] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 355–369, 2007.

[9] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta path-based top-k similarity search in heterogeneous information networks," *Vldb 2011*, vol. 3, no. 2, pp. 1–12, 2011.

[10] Y. Sun, B. Norick, and J. Han, "PathSelClus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," *ACMTrans. Knowl. Discov. Data*, vol. 7, no. 3, 2013.

[11] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, vol. 2, no. 8, pp. 0696–0701, 2005.

[12] J. Cohen, "Using effect size or why the p value is not enough," *Journal of Graduate Medical Education*, no. September, pp. 279–282, 2012.

[13] D. B. West, *Introduction to Graph Theory*, 2nd ed. Pearson, 2000.

[14] N. Biggs, *Algebraic Graph Theory*, 2nd ed. Cambridge University Press, 1993.

[15] J. A. Blake et al., "Gene ontology consortium: Going forward," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

[16] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[17] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin, "Pfam: Multiple sequence alignments and HMM-profiles of protein domains," *Nucleic Acids Research*, no. 1, pp. 320–322. [Online]. Available: https://academic.oup.com/nar/article/26/1/320/2379329

[18] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool." *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–10, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283605803602

[19] L. Salwinski, "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 449D–451, 2004. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh086

[20] D. Szklarczyk et al., "The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible," *Nucleic Acids Research*, vol. 45, no. D1, pp. D362–D368, 2017.

[21] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the National Academy of Sciences USA*, vol. 97, no. 21, pp. 11 149–11 152, 2000. [Online]. Available: http://www.pnas.org/content/97/21/11149.full

[22] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005. [Online]. Available: http://jcs.biologists.org/cgi/doi/10.1242/jcs.02714

[23] P. A. Gagniuc, "From Observation to Simulation," in *Markov Chains: From Theory to Implementation and Experimentation.* John Wiley & Sons, Inc., 2017, ch. 2.3, pp. 9–14.

[24] G. R. Norman and D. L. Streiner, "The Normal Distribution," in *Biostatistics: The Bare Essentials*, 3rd ed. BC Decker Inc, 2008, ch. 4, pp. 32–36.

[25] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[27] J. L. Devore, *Probability & Statistics for Engineering and the Sciences*, 8th ed. Cengage Learning, 2012.

[28] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pp. 233–240, 2006. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1143844.1143874

[29] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[30] K. A. Tryka et al., "NCBI's database of genotypes and phenotypes: DbGaP," *Nucleic Acids Research*, vol. 42, no. D1, pp. 975–979, 2014.

[31] G. S. Cowley et al., "Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies," *Scientific Data*, vol. 1, pp. 1–12, 2014.

[32] S. M. Sunkin, L. Ng, C. Lau, T. Dolbeare, T. L. Gilbert, C. L. Thompson, M. Hawrylycz, and C. Dang, "Allen brain atlas: An integrated spatio-temporal portal for exploring the central nervous system," *Nucleic Acids Research*, vol. 41, no. D1, 2013.

[33] M. V. Kuleshov et al., "Enrichr: A comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.

[34] H. Xu, C. Baroukh, R. Dannenfelser, E. Y. Chen, C. M. Tan, Y. Kou, Y. E. Kim, I. R. Lemischka, and A. Ma'ayan, "ESCAPE: Database for integrating high-content published data collected from human and mouse embryonic stem cells," *Database*, vol. 2013, pp. 1–12, 2013.

[35] A. C. Culhane et al., "GeneSigDB: A manually curated database and resource for analysis of gene expression signatures," *Nucleic Acids Research*, vol. 40, no. D1, pp. 1060–1066, 2012.

[36] T. Barrett et al., "NCBI GEO: Archive for functional genomics data sets - Update," *Nucleic Acids Research*, vol. 41, no. D1, pp. 991–995, 2013.

[37] A. B. Keenan et al., "The library of integrated network-based cellular signatures NIH program: System-level cataloging of human cells response to perturbations," *Cell Systems*, vol. 6, no. 1, pp. 13–24, 2018.

[38] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, "Pathway commons, a web resource for biological pathway data," *Nucleic Acids Research*, vol. 39, no. SUPPL. 1, pp. 685–690, 2011.

[39] A. Fabregat et al., "The reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 46, no. D1, pp. D649–D655, 2018.

[40] V. Agarwal, G. W. Bell, J. W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *eLife*, vol. 4, no. 8, pp. 1–38, Aug. 2015.

[41] C. Blatti and S. Sinha, "Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks," *Bioinformatics*, vol. 32, no. 14, pp. 2167–2175, 2016.

[42] M. P. Goetz et al., "Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer," *Journal of the National Cancer Institute*, vol. 109, no. 7, pp. 1–9, 2017.

[43] S. Tabariès and P. M. Siegel, "The role of claudins in cancer metastasis," *Oncogene*, vol. 36, no. 9, pp. 1176–1190, 2017.

[44] A. Prat, J. S. Parker, O. Karginova, C. Fan, C. Livasy, J. I. Herschkowitz, X. He, and C. M. Perou, "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer," *Breast Cancer Research*, vol. 12, no. 5, 2010.

[45] K. Dias, A. Dvorkin-Gheva, R. M. Hallett, Y. Wu, J. Hassell, G. R. Pond, M. Levine, T. Whelan, and A. L. Bane, "Claudin-low breast cancer; clinical & pathological characteristics," *PLoS ONE*, vol. 12, no. 1, pp. 1–17, 2017.

[46] C. A. Borgoño and E. P. Diamandis, "The emerging roles of human tissue kallikreins in cancer," *Nature Reviews Cancer*, vol. 4, no. 11, pp. 876–890, 2004.

[47] M. Paliouras, C. Borgono, and E. P. Diamandis, "Human tissue kallikreins: The cancer biomarker family," *Cancer Letters*, vol. 249, no. 1, pp. 61–79, 2007.

[48] P. Chen, M. Cescon, and P. Bonaldo, "Collagen VI in cancer and its biological mechanisms," *Trends in Molecular Medicine*, vol. 19, no. 7, pp. 410–417, 2013.

[49] S. H. Barsky, C. N. Rao, G. R. Grotendorst, and L. A. Liotta, "Increased content of Type V Collagen in desmoplasia of human breast carcinoma." *The American journal of pathology*, vol. 108, no. 3, pp. 276–83, 1982.

[50] G. Huang, G. Ge, V. Izzi, and D. S. Greenspan, "α3 Chains of type v collagen regulate breast tumour growth via glypican-1," *Nature Communications*, vol. 8, pp. 1–17, 2017. [Online]. Available: http://dx.doi.org/10.1038/ncomms14351

[51] X. Solé et al., "Discovery and validation of new potential biomarkers for early detection of colon cancer," *PLoS ONE*, vol. 9, no. 9, 2014.

[52] J. Pathol, F. Stenback, J. Risteli, a. Jukkola, and L. Risteli, "Aberrant type I and type III collagen gene expression in human breast cancer in vivo," *Journal of Pathology*, vol. 268, no. July, pp. 262–268, 1998.

[53] J. Freire, S. Domínguez-Hormaetxe, S. Pereda, A. De Juan, A. Vega, L. Simón, and J. Gómez-Román, "Collagen, type XI, alpha 1: An accurate marker for differential diagnosis of breast carcinoma invasiveness in core needle biopsies," *Pathology Research and Practice*, vol. 210, no. 12, pp. 879–884, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.prp.2014.07.012

[54] Z. Raglow and S. M. Thomas, "Tumor matrix protein collagen XIα1 in cancer," *Cancer Letters*, vol. 357, no. 2, pp. 448–453, 2015.

[55] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "Revigo summarizes and visualizes long lists of gene ontology terms," *PLoS ONE*, vol. 6, no. 7, 2011.

[56] D. Warde-Farley et al., "The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. SUPPL. 2, pp. 214–220, 2010.

[57] A. Hyvärinen, J. Karhunen, and E. Oja, "What is Independent Component Analysis?" in *Independent Component Analysis*, 1st ed., S. Haykin, Ed. Wiley-Interscience, 2001, ch. 7, p. 504.

[58] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

APPENDIX A

SUMMARY OF ALL TESTED GENE SETS

Table A.1 contains an entry for every gene set tested in Chapter 5. For each set, the collection of which it was a part and the number of genes in the set are indicated. The next three columns relate to the node degree for genes in the set. The median degree for the set is followed by the proportional change to the median degree for the top-N genes ranked by GeneSet MAPR. For example, if the original set contained 50 genes with a median degree of 100, and the top 50 genes returned by MAPR had a median degree of 210, then a value of 2.1 indicates the proportional change in the median from 100 to 210. The column Δ200+ indicates the difference in proportion of genes with a degree of 200 or greater. For example, if a set of 100 genes contained 20 with degree at or above 200 (a proportion of 0.2), and the top 100 MAPR genes contained 30 (a proportion of 0.3), then a value of positive 0.1 indicates this difference. The final three columns show the AUC achieved by MAPR, DRaWR, and LASSO ensemble, respectively.

Table A.1: All Tested Gene Sets with Node Degree Statistics and Method AUC Values

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Achilles	22RV1 PROSTATE	141	119	5.4	0.65	0.679	0.559	0.561
Achilles	697 HAEMATOPOIETIC AND LYMPHOID TISSUE	294	107	1.4	0.33	0.685	0.637	0.538
Achilles	786O KIDNEY	167	153	4.5	0.52	0.726	0.609	0.689
Achilles	A1207 CENTRAL NERVOUS SYSTEM	295	132	1.5	0.37	0.711	0.608	0.581
Achilles	A172 CENTRAL NERVOUS SYSTEM	171	122	4.7	0.52	0.704	0.571	0.648
Achilles	A204 SOFT TISSUE	182	156	2.1	0.46	0.775	0.633	0.739
Achilles	A2058 SKIN	222	142	2.4	0.45	0.759	0.667	0.692
Achilles	A549 LUNG	152	131	4.6	0.51	0.694	0.542	0.647
Achilles	A673 BONE	194	112	1.4	0.42	0.695	0.594	0.594
Achilles	ACHN KIDNEY	234	134	2.0	0.45	0.743	0.612	0.643
Achilles	AGS STOMACH	439	130	1.8	0.41	0.731	0.693	0.703
Achilles	AM38 CENTRAL NERVOUS SYSTEM	191	143	4.0	0.61	0.706	0.594	0.656

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Achilles	AML193 HAEMATOPOIETIC AND LYMPHOID TISSUE	222	129	4.2	0.65	0.702	0.559	0.663
Achilles	ASPC1 PANCREAS	191	121	1.2	0.31	0.720	0.633	0.648
Achilles	BT20 BREAST	175	137	2.5	0.39	0.728	0.658	0.695
Achilles	BT474 BREAST	304	137	2.5	0.53	0.731	0.629	0.670
Achilles	BXPC3 PANCREAS	135	131	5.4	0.64	0.748	0.606	0.712
Achilles	C2BBE1 LARGE INTESTINE	189	115	2.9	0.37	0.681	0.543	0.614
Achilles	C32 SKIN	254	127	0.8	0.24	0.703	0.620	0.627
Achilles	CADOES1 BONE	228	116	1.2	0.29	0.689	0.606	0.521
Achilles	CAL120 BREAST	171	151	3.6	0.57	0.757	0.658	0.727
Achilles	CAL51 BREAST	199	100	1.4	0.33	0.681	0.594	0.552
Achilles	CALU1 LUNG	295	123	2.8	0.58	0.677	0.654	0.532
Achilles	CAOV3 OVARY	148	143	2.8	0.36	0.750	0.606	0.694
Achilles	CAOV4 OVARY	242	133	0.9	0.23	0.754	0.680	0.683
Achilles	CAS1 CENTRAL NERVOUS SYSTEM	396	122	3.0	0.60	0.701	0.583	0.594
Achilles	CFPAC1 PANCREAS	263	120	2.7	0.54	0.723	0.548	0.572
Achilles	CH157MN CENTRAL NERVOUS SYSTEM	194	128	3.8	0.59	0.711	0.652	0.694
Achilles	COLO205 LARGE INTESTINE	238	132	3.1	0.50	0.708	0.579	0.617
Achilles	COLO704 OVARY	247	123	1.5	0.44	0.724	0.698	0.675
Achilles	COLO741 SKIN	111	125	6.8	0.59	0.676	0.625	0.634
Achilles	COLO783 SKIN	292	137	1.8	0.38	0.709	0.605	0.584
Achilles	CORL23 LUNG	365	97	0.6	0.18	0.671	0.631	0.513
Achilles	COV362 OVARY	435	132	1.8	0.44	0.734	0.668	0.700
Achilles	COV434 OVARY	137	134	3.2	0.51	0.693	0.536	0.645
Achilles	COV504 OVARY	195	121	2.9	0.43	0.685	0.580	0.627
Achilles	COV644 OVARY	169	131	1.2	0.38	0.730	0.637	0.689
Achilles	DBTRG05MG CENTRAL NERVOUS SYSTEM	158	122	0.6	0.17	0.699	0.599	0.624
Achilles	DKMG CENTRAL NERVOUS SYSTEM	144	133	2.2	0.36	0.683	0.594	0.608
Achilles	DLD1 LARGE INTESTINE	339	116	1.1	0.32	0.699	0.666	0.582
Achilles	EFE184 ENDOMETRIUM	296	124	1.1	0.35	0.726	0.692	0.682
Achilles	EFM19 BREAST	427	133	1.6	0.38	0.748	0.671	0.702
Achilles	EFO21 OVARY	490	134	1.8	0.45	0.738	0.662	0.691
Achilles	EFO27 OVARY	76	116	1.1	0.29	0.700	0.574	0.732
Achilles	EW8 BONE	173	125	2.8	0.53	0.690	0.527	0.641
Achilles	EWS502 BONE	308	104	0.5	0.19	0.681	0.609	0.574
Achilles	F36P HAEMATOPOIETIC AND LYMPHOID TISSUE	268	121	2.8	0.51	0.691	0.636	0.675
Achilles	GB1 CENTRAL NERVOUS SYSTEM	138	131	2.7	0.50	0.702	0.647	0.624
Achilles	GP2D LARGE INTESTINE	333	129	3.0	0.59	0.715	0.611	0.644
Achilles	HCC1187 BREAST	342	125	3.3	0.64	0.704	0.574	0.613
Achilles	HCC1395 BREAST	126	128	4.4	0.36	0.719	0.605	0.671
Achilles	HCC1954 BREAST	343	134	2.5	0.54	0.728	0.631	0.672
Achilles	HCC2218 BREAST	324	146	2.3	0.47	0.731	0.685	0.673
Achilles	HCC2814 LUNG	229	136	2.6	0.46	0.707	0.639	0.656
Achilles	HCC364 LUNG	236	125	3.4	0.48	0.727	0.581	0.628
Achilles	HCC44 LUNG	189	126	3.8	0.60	0.729	0.551	0.648
Achilles	HCC70 BREAST	327	165	1.5	0.33	0.786	0.710	0.766
Achilles	HCC827 LUNG	201	127	3.8	0.54	0.705	0.613	0.603
Achilles	HCC827GR5 LUNG	119	138	2.7	0.57	0.735	0.619	0.670
Achilles	HCT116 LARGE INTESTINE	369	147	2.3	0.54	0.772	0.679	0.732
Achilles	HEC1A ENDOMETRIUM	170	118	0.6	0.21	0.693	0.549	0.598
Achilles	HEYA8 OVARY	133	111	1.3	0.35	0.705	0.653	0.656
Achilles	HL60 HAEMATOPOIETIC AND LYMPHOID TISSUE	541	114	2.1	0.41	0.672	0.591	0.600
Achilles	HLF LIVER	99	134	0.6	0.31	0.719	0.629	0.710
Achilles	HNT34 HAEMATOPOIETIC AND LYMPHOID TISSUE	263	120	3.8	0.56	0.674	0.566	0.580
Achilles	HPAC PANCREAS	142	125	4.2	0.39	0.763	0.686	0.718
Achilles	HPAFII PANCREAS	244	125	3.0	0.59	0.709	0.594	0.654

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Achilles	HS683 CENTRAL NERVOUS SYSTEM	244	127	3.5	0.63	0.728	0.660	0.652
Achilles	HS766T PANCREAS	112	109	5.7	0.46	0.706	0.572	0.669
Achilles	HS944T SKIN	226	126	2.6	0.56	0.657	0.594	0.569
Achilles	HT1197 URINARY TRACT	189	138	2.4	0.34	0.728	0.638	0.654
Achilles	HT29 LARGE INTESTINE	366	143	1.8	0.42	0.777	0.714	0.747
Achilles	HT55 LARGE INTESTINE	470	133	2.9	0.60	0.706	0.634	0.655
Achilles	HUG1N STOMACH	282	128	3.2	0.63	0.700	0.604	0.607
Achilles	HUTU80 SMALL INTESTINE	182	120	1.2	0.33	0.712	0.643	0.653
Allen Brain	DN AMYGDALOHIPPOCAMPAL AREA	295	77	0.0	−0.01	0.673	0.603	0.518
Allen Brain	DN INTERMEDIATE PART OF R3B	295	96	0.3	−0.07	0.664	0.594	0.507
Allen Brain	DN INTERNAL GRANULAR LAYER OF CBVCX	298	102	2.4	0.48	0.738	0.656	0.644
Allen Brain	DN LATEROSTRIATAL STRIPE	295	71	0.2	0.08	0.551	0.586	0.489
Allen Brain	DN LAYER 2 OF OCX	295	97	0.4	0.15	0.715	0.640	0.721
Allen Brain	DN M2 PART OF PARARUBRAL NUCLEUS	295	76	0.0	0.02	0.603	0.546	0.520
Allen Brain	DN MANTLE ZONE OF PHYB P	293	81	0.5	0.12	0.629	0.535	0.505
Allen Brain	DN MANTLE ZONE OF R5BL	294	94	0.1	0.00	0.676	0.607	0.468
Allen Brain	DN MANTLE ZONE OF TPAA	296	93	3.9	0.53	0.688	0.608	0.611
Allen Brain	DN R5 PART OF PARVOCELLULAR MEDIAL VESTIBULAR NUCLEUS	295	71	1.2	0.21	0.623	0.552	0.492
Allen Brain	DN R7 PART OF SPINAL VESTIBULAR NUCLEUS	296	89	4.0	0.55	0.689	0.601	0.519
Allen Brain	DN SEPTODIAGONAL TRANSITION AREA	297	89	0.6	0.21	0.652	0.571	0.549
Allen Brain	DN SUBICULUM VENTRAL PART MOLECULAR LAYER	294	74	0.0	−0.06	0.660	0.578	0.536
Allen Brain	DN ZONA INCERTA COMPLEX	293	75	1.4	0.30	0.570	0.520	0.469
Allen Brain	UP AGRANULAR INSULAR AREA DORSAL PART LAYER 6A	298	87	0.1	0.05	0.667	0.548	0.563
Allen Brain	UP ANTERIOR PART OF INT	290	66	0.2	−0.02	0.575	0.559	0.499
Allen Brain	UP BED NUCLEI OF THE STRIA TERMINALIS ANTERIOR DIVISION J	295	96	0.0	−0.02	0.712	0.678	0.537
Allen Brain	UP COLLICULAR ROSTRAL MIDBRAIN TECTUM	297	98	2.1	0.33	0.615	0.546	0.565
Allen Brain	UP DORSAL ENTOPEDUNCULAR NUCLEUS	296	92	−0.3	−0.03	0.646	0.527	0.537
Allen Brain	UP DORSAL PEDUNCULAR AREA LAYER 5	296	89	3.4	0.57	0.639	0.553	0.551
Allen Brain	UP FIELD CA1 STRATUM ORIENS	336	96	1.5	0.31	0.696	0.579	0.532
Allen Brain	UP INTERMEDIATE STRATUM OF R9TR	294	81	2.7	0.43	0.608	0.598	0.497
Allen Brain	UP ISTHMIC LIMINAL PART OF THE PERIAQUEDUCTAL GRAY	297	78	0.2	0.20	0.597	0.567	0.490
Allen Brain	UP ISTHMIC PART OF BASOLATERAL ISTHMIC RETICULAR FORMATION	298	80	0.1	0.00	0.658	0.585	0.469
Allen Brain	UP LATERAL DORSAL NUCLEUS OF THALAMUS	291	83	1.3	0.29	0.667	0.618	0.539
Allen Brain	UP LIMINAL ALAR DOMAIN OF M2	294	79	0.1	0.03	0.619	0.584	0.510
Allen Brain	UP MANTLE ZONE OF R6VE	296	70	−0.2	−0.10	0.634	0.566	0.510

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Allen Brain	UP MEDIAL AMYGDALA POSTERODORSAL PART	298	99	2.4	0.39	0.640	0.601	0.493
Allen Brain	UP ORBITAL AREA VENTROLATERAL PART LAYER 2 3	296	92	1.7	0.30	0.630	0.599	0.522
Allen Brain	UP PERIVENTRICULAR STRATUM OF M1B	299	87	0.0	−0.02	0.660	0.626	0.528
Allen Brain	UP PERIVENTRICULAR STRATUM OF R9LIM	295	85	0.0	−0.02	0.646	0.556	0.500
Allen Brain	UP PONTINE RETICULAR NUCLEUS CAUDAL PART	300	73	−0.2	−0.04	0.609	0.560	0.491
Allen Brain	UP PRIMARY SOMATOSENSORY AREA TRUNK LAYER 4	293	86	0.9	0.20	0.615	0.563	0.519
Allen Brain	UP R2 PART OF BASOINTERMEDIATE RETICULAR FORMATION	299	70	0.8	0.15	0.650	0.556	0.529
Allen Brain	UP R2 PART OF THE VENTRAL PARVICELLULAR RETICULAR FORMATION	297	76	0.2	0.02	0.614	0.576	0.510
Allen Brain	UP R5 PART OF POSTEROVENTRAL COCHLEAR NUCLEUS	290	93	1.4	0.33	0.640	0.520	0.567
Allen Brain	UP R9 PART OF THE VESTIBULAR COLUMN	301	88	0.7	0.10	0.591	0.604	0.469
Allen Brain	UP SUBPALLIAL SEPTUM	295	84	−0.1	−0.02	0.692	0.604	0.501
Allen Brain	UP SUPERFICIAL STRATUM OF CEREBELLAR HEMISPHERE	294	95	0.1	0.02	0.680	0.598	0.551
Allen Brain	UP SUPERFICIAL STRATUM OF TG	298	84	−0.2	−0.10	0.643	0.624	0.552
dbGaP	Alcoholism-93	93	90	0.1	0.03	0.692	0.616	0.649
dbGaP	Alzheimer Disease-79	79	101	0.2	0.24	0.678	0.572	0.612
dbGaP	Amyotrophic Lateral Sclerosis-76	76	87	0.1	0.04	0.675	0.595	0.654
dbGaP	Arteries-88	88	94	0.5	0.30	0.667	0.643	0.604
dbGaP	Asthma-106	106	122	2.7	0.48	0.716	0.565	0.696
dbGaP	Attention Deficit Disorder with Hyperactivity-99	99	90	−0.1	0.26	0.676	0.567	0.615
dbGaP	Bipolar Disorder-102	102	75	0.1	0.05	0.656	0.710	0.614
dbGaP	Blood Pressure-447	447	90	0.4	0.10	0.689	0.642	0.532
dbGaP	Body Height-379	379	100	0.7	0.23	0.687	0.648	0.619
dbGaP	Body Mass Index-434	434	85	0.4	0.05	0.699	0.694	0.575
dbGaP	Body Weight-212	212	88	2.0	0.36	0.681	0.628	0.563
dbGaP	Body Weights and Measures-85	85	95	2.4	0.37	0.653	0.488	0.587
dbGaP	Bone Density-73	73	113	0.3	0.15	0.667	0.586	0.637
dbGaP	C Reactive Protein-87	87	93	−0.1	−0.02	0.681	0.619	0.595
dbGaP	Cholesterol HDL-354	354	88	0.1	−0.02	0.659	0.648	0.488
dbGaP	Cholesterol LDL-301	301	97	0.1	0.03	0.696	0.604	0.519
dbGaP	Cholesterol-267	267	95	0.4	0.17	0.699	0.621	0.507
dbGaP	Coronary Artery Disease-201	201	91	3.5	0.52	0.688	0.611	0.521
dbGaP	Coronary Disease-81	81	94	0.3	0.10	0.675	0.595	0.649
dbGaP	Creatinine-91	91	102	0.1	0.11	0.712	0.669	0.618
dbGaP	Diabetes Mellitus Type 1-74	74	99	3.4	0.50	0.576	0.693	0.601
dbGaP	Diabetes Mellitus Type 2-84	84	99	3.1	0.33	0.721	0.662	0.701
dbGaP	Diabetes Mellitus-69	69	87	−0.2	0.03	0.720	0.658	0.694
dbGaP	Echocardiography-271	271	87	0.1	0.02	0.675	0.670	0.523
dbGaP	Elbow-71	71	95	0.3	0.09	0.673	0.451	0.612
dbGaP	Electrocardiography-84	84	114	0.1	0.04	0.716	0.685	0.667
dbGaP	Erythrocyte Count-114	114	91	0.2	0.16	0.699	0.609	0.598
dbGaP	Fibrinogen-84	84	122	1.3	0.36	0.707	0.635	0.655
dbGaP	Glucose-144	144	72	0.0	−0.03	0.691	0.614	0.594
dbGaP	Heart Failure-182	182	82	0.3	0.13	0.637	0.604	0.563
dbGaP	Heart Rate-155	155	91	0.1	0.17	0.681	0.657	0.526
dbGaP	Hemoglobin A Glycosylated-125	125	96	3.3	0.35	0.630	0.602	0.588

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
dbGaP	Hip-198	198	92	0.4	0.12	0.672	0.671	0.515
dbGaP	Insulin-87	87	124	2.0	0.38	0.713	0.623	0.633
dbGaP	Iron-152	152	87	−0.2	−0.08	0.720	0.660	0.565
dbGaP	Lipids-85	85	83	0.2	0.01	0.672	0.664	0.608
dbGaP	Lipoproteins VLDL-84	84	104	2.8	0.56	0.686	0.693	0.572
dbGaP	Lipoproteins-79	79	82	−0.1	0.05	0.740	0.699	0.657
dbGaP	Lupus Erythematosus Systemic-86	86	107	4.4	0.44	0.622	0.546	0.633
dbGaP	Macular Degeneration-118	118	76	1.5	0.29	0.591	0.525	0.578
dbGaP	Multiple Sclerosis-82	82	111	3.8	0.51	0.617	0.660	0.609
dbGaP	Myocardial Infarction-228	228	86	0.2	−0.06	0.696	0.652	0.524
dbGaP	Neuroblastoma-100	100	84	−0.2	−0.06	0.624	0.511	0.591
dbGaP	Parkinson Disease-130	130	76	5.8	0.41	0.664	0.628	0.579
dbGaP	Respiratory Function Tests-148	148	89	0.1	0.03	0.719	0.695	0.589
dbGaP	Schizophrenia-82	82	90	0.6	0.24	0.666	0.575	0.577
dbGaP	Stroke-284	284	83	0.1	−0.01	0.667	0.613	0.510
dbGaP	Triglycerides-240	240	85	0.1	−0.03	0.673	0.643	0.497
dbGaP	Tunica Media-103	103	100	−0.3	−0.01	0.704	0.681	0.631
dbGaP	Waist Circumference-144	144	94	−0.1	−0.04	0.673	0.623	0.581
dbGaP	Waist Hip Ratio-92	92	76	2.2	0.36	0.647	0.694	0.612
Enrichr Path	NCI DIRECT P53 EFFECTORS HOMO SAPIENS 67C3B75D 6191 11E5 8AC5 0	135	218	0.0	−0.01	0.936	0.840	0.870
Enrichr Path	WIKIPATH ADIPOGENESIS GENES MUS MUSCULUS WP447	128	300	−0.6	−0.27	0.944	0.965	0.912
Enrichr Path	WIKIPATH ADIPOGENESIS HOMO SAPIENS WP236	129	302	−0.5	−0.26	0.951	0.970	0.926
Enrichr Path	WIKIPATH ALZHEIMERS DISEASE HOMO SAPIENS WP2059	120	193	−0.4	−0.28	0.899	0.988	0.893
Enrichr Path	WIKIPATH BDNF SIGNALING PATHWAY HOMO SAPIENS WP2380	143	453	0.0	0.02	0.969	0.974	0.937
Enrichr Path	WIKIPATH CALCIUM REGULATION IN THE CARDIAC CELL HOMO SAPIENS WP	149	156	−0.1	−0.10	0.989	0.978	0.970
Enrichr Path	WIKIPATH CALCIUM REGULATION IN THE CARDIAC CELL MUS MUSCULUS WP	150	153	−0.1	−0.05	0.990	0.980	0.984
Enrichr Path	WIKIPATH CELL CYCLE HOMO SAPIENS WP179	103	324	−0.1	−0.12	0.972	0.970	0.955
Enrichr Path	WIKIPATH CHEMOKINE SIGNALING PATHWAY MUS MUSCULUS WP2292	164	244	−0.1	−0.06	0.991	0.995	0.982
Enrichr Path	WIKIPATH CIRCADIAN RYTHM RELATED GENES HOMO SAPIENS WP3594	200	228	1.3	0.18	0.928	0.935	0.977
Enrichr Path	WIKIPATH EGFR1 SIGNALING PATHWAY MUS MUSCULUS WP572	171	366	0.1	−0.05	0.969	0.907	0.939
Enrichr Path	WIKIPATH ELECTRON TRANSPORT CHAIN HOMO SAPIENS WP111	103	127	0.0	−0.02	0.991	0.999	0.990
Enrichr Path	WIKIPATH ESC PLURIPOTENCY PATHWAYS MUS MUSCULUS WP339	116	335	−0.4	−0.17	0.992	0.993	0.978
Enrichr Path	WIKIPATH FOCAL ADHESION HOMO SAPIENS WP306	187	313	0.1	−0.01	0.991	0.992	0.988
Enrichr Path	WIKIPATH FOCAL ADHESION MUS MUSCULUS WP85	182	293	0.2	0.00	0.986	0.982	0.980

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Enrichr Path	WIKIPATH GPCRS CLASS A RHODOPSIN LIKE HOMO SAPIENS WP455	258	202	−0.1	−0.08	0.989	0.992	0.953
Enrichr Path	WIKIPATH GPCRS CLASS A RHODOPSIN LIKE MUS MUSCULUS WP189	171	200	−0.1	−0.13	0.989	0.994	0.964
Enrichr Path	WIKIPATH INSULIN SIGNALING HOMO SAPIENS WP481	160	384	0.2	0.04	0.976	0.980	0.957
Enrichr Path	WIKIPATH INTEGRATED BREAST CANCER PATHWAY HOMO SAPIENS WP1984	152	368	0.4	0.11	0.917	0.906	0.876
Enrichr Path	WIKIPATH MAPK SIGNALING PATHWAY MUS MUSCULUS WP493	155	378	0.0	0.00	0.982	0.972	0.981
Enrichr Path	WIKIPATH METAPATHWAY BIOTRANSFORMATION HOMO SAPIENS WP702	176	76	0.1	−0.02	0.994	0.997	0.992
Enrichr Path	WIKIPATH MRNA PROCESSING HOMO SAPIENS WP411	126	187	0.0	−0.06	0.989	0.975	0.967
Enrichr Path	WIKIPATH MYOMETRIAL RELAXATION AND CONTRACTION PATHWAYS HOMO SA	155	194	0.4	0.12	0.976	0.957	0.955
Enrichr Path	WIKIPATH MYOMETRIAL RELAXATION AND CONTRACTION PATHWAYS MUS MUS	151	191	1.3	0.16	0.974	0.974	0.965
Enrichr Path	WIKIPATH NEURAL CREST DIFFERENTIATION HOMO SAPIENS WP2064	101	311	−0.5	−0.29	0.972	0.977	0.947
Enrichr Path	WIKIPATH NRF2 PATHWAY HOMO SAPIENS WP2884	145	90	1.9	0.32	0.970	0.937	0.966
Enrichr Path	WIKIPATH ODORANT GPCRS MUS MUSCULUS WP1397	109	96	0.6	0.05	0.977	0.983	0.970
Enrichr Path	WIKIPATH PLURINETWORK MUS MUSCULUS WP1763	283	303	0.2	0.01	0.940	0.929	0.888
Enrichr Path	WIKIPATH PODNET PROTEIN PROTEIN INTERACTIONS IN THE PODOCYTE M	303	195	0.5	0.11	0.922	0.882	0.885
Enrichr Path	WIKIPATH PURINE METABOLISM MUS MUSCULUS WP2185	158	107	0.1	0.05	0.997	0.998	0.993
Enrichr Path	WIKIPATH REGULATION OF ACTIN CYTOSKELETON HOMO SAPIENS WP51	148	248	0.0	−0.01	0.987	0.993	0.986
Enrichr Path	WIKIPATH SENESCENCE AND AUTOPHAGY IN CANCER HOMO SAPIENS WP615	105	287	0.6	−0.02	0.947	0.978	0.918
Enrichr Path	WIKIPATH SIDS SUSCEPTIBILITY PATHWAYS HOMO SAPIENS WP706	159	249	−0.1	−0.06	0.939	0.918	0.917
Enrichr Path	WIKIPATH SPINAL CORD INJURY HOMO SAPIENS WP2431	118	266	−0.5	−0.25	0.925	0.916	0.891
Enrichr Path	WIKIPATH TGF BETA SIGNALING PATHWAY HOMO SAPIENS WP366	132	438	0.1	−0.10	0.973	0.967	0.947

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Enrichr Path	WIKIPATH TNF ALPHA NF KB SIGNALING PATHWAY MUS MUSCULUS WP246	178	315	0.0	−0.07	0.962	0.946	0.922
Enrichr Path	WIKIPATH TOLL LIKE RECEPTOR SIGNALING PATHWAY HOMO SAPIENS WP75	102	353	−0.3	−0.09	0.996	0.999	0.997
Enrichr Path	WIKIPATH WNT SIGNALING PATHWAY AND PLURIPOTENCY HOMO SAPIENS WP	101	302	0.4	0.04	0.977	0.961	0.966
Enrichr Path	WIKIPATH WNT SIGNALING PATHWAY NETPATH MUS MUSCULUS WP539	106	428	0.1	−0.01	0.977	0.981	0.971
Enrichr Path	WIKIPATH XPODNET PROTEIN PROTEIN INTERACTIONS IN THE PODOCYTE	802	193	0.3	0.11	0.909	0.889	0.852
Enrichr Pheno	DBGAP BODY MASS INDEX	435	84	0.4	0.00	0.695	0.671	0.570
Enrichr Pheno	DBGAP CHOLESTEROL	265	96	1.1	0.29	0.648	0.541	0.506
Enrichr Pheno	DBGAP CHOLESTEROL HDL	351	88	0.0	−0.03	0.685	0.670	0.497
Enrichr Pheno	DBGAP CHOLESTEROL LDL	299	98	0.1	0.04	0.684	0.626	0.523
Enrichr Pheno	DBGAP CORONARY ARTERY DISEASE	201	91	3.5	0.49	0.675	0.610	0.502
Enrichr Pheno	DBGAP ECHOCARDIOGRAPHY	268	88	0.1	0.00	0.679	0.665	0.512
Enrichr Pheno	DBGAP HEMOGLOBIN A GLYCOSYLATED	126	96	4.7	0.52	0.620	0.609	0.595
Enrichr Pheno	DBGAP HIP	198	92	0.7	0.18	0.656	0.677	0.595
Enrichr Pheno	DBGAP MACULAR DEGENERATION	116	76	1.4	0.30	0.611	0.667	0.558
Enrichr Pheno	DBGAP NEUROBLASTOMA	101	84	−0.3	−0.07	0.632	0.511	0.612
Enrichr Pheno	DBGAP TUNICA MEDIA	104	100	0.2	0.19	0.687	0.637	0.598
Enrichr Pheno	DBGAP WAIST CIRCUMFERENCE	143	94	−0.1	0.03	0.658	0.642	0.602
Enrichr Pheno	HPO ABDOMINAL PAIN HP 0002027	118	162	−0.3	−0.04	0.876	0.840	0.804
Enrichr Pheno	HPO AUTOSOMAL RECESSIVE INHERITANCE HP 0000007	1,706	108	0.0	0.00	0.848	0.819	0.775
Enrichr Pheno	HPO BLEPHAROPHIMOSIS HP 0000581	103	161	2.9	0.23	0.827	0.802	0.771
Enrichr Pheno	HPO CAMPTODACTYLY OF FINGER HP 0100490	121	126	1.2	0.21	0.872	0.836	0.765
Enrichr Pheno	HPO CEREBRAL CORTICAL ATROPHY HP 0002120	152	125	0.1	0.13	0.787	0.762	0.724
Enrichr Pheno	HPO ELEVATED HEPATIC TRANSAMINASES HP 0002910	111	105	−0.2	−0.05	0.866	0.837	0.823
Enrichr Pheno	HPO EPICANTHUS HP 0000286	232	136	1.2	0.23	0.849	0.791	0.726
Enrichr Pheno	HPO FRONTAL BOSSING HP 0002007	210	141	0.1	0.08	0.850	0.808	0.811
Enrichr Pheno	HPO HEPATOMEGALY HP 0002240	329	115	0.0	0.03	0.864	0.844	0.783
Enrichr Pheno	HPO HETEROGENEOUS HP 0001425	148	136	−0.1	−0.05	0.880	0.874	0.779
Enrichr Pheno	HPO HYPOPLASIA OF THE CORPUS CALLOSUM HP 0002079	102	109	−0.6	−0.15	0.791	0.770	0.735
Enrichr Pheno	HPO MALAR FLATTENING HP 0000272	185	139	1.4	0.20	0.846	0.771	0.778
Enrichr Pheno	HPO MULTICYSTIC KIDNEY DYSPLASIA HP 0000003	100	134	6.6	0.47	0.890	0.903	0.885

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Enrichr Pheno	HPO OPACIFICATION OF THE CORNEAL STROMA HP 0007759	135	104	−0.3	0.04	0.877	0.812	0.832
Enrichr Pheno	HPO OPTIC ATROPHY HP 0000648	360	109	0.4	0.16	0.862	0.823	0.771
Enrichr Pheno	HPO PATENT DUCTUS ARTERIOSUS HP 0001643	124	161	−0.2	−0.04	0.855	0.807	0.814
Enrichr Pheno	HPO PECTUS EXCAVATUM HP 0000767	132	171	−0.5	−0.20	0.828	0.768	0.787
Enrichr Pheno	HPO PES PLANUS HP 0001763	105	162	0.1	0.08	0.857	0.865	0.838
Enrichr Pheno	HPO PHOTOPHOBIA HP 0000613	147	111	0.5	0.23	0.884	0.866	0.839
Enrichr Pheno	HPO PROGRESSIVE DISORDER HP 0003676	142	115	−0.1	0.05	0.861	0.874	0.791
Enrichr Pheno	HPO PROTEINURIA HP 0000093	105	130	3.8	0.31	0.838	0.835	0.781
Enrichr Pheno	HPO SKIN ULCER HP 0200042	118	173	1.4	0.26	0.849	0.801	0.785
Enrichr Pheno	HPO UPSLANTED PALPEBRAL FISSURE HP 0000582	107	135	6.7	0.53	0.846	0.841	0.810
Enrichr Pheno	HPO VOMITING HP 0002013	108	108	0.5	0.18	0.910	0.863	0.892
Enrichr Pheno	OMIM EXP CARDIOMYOPATHY	111	280	0.7	0.08	0.949	0.947	0.919
Enrichr Pheno	OMIM EXP CARDIOMYOPATHY DILATED	104	288	−0.1	−0.10	0.948	0.953	0.920
Enrichr Pheno	OMIM EXP COLORECTAL CANCER	102	334	0.4	0.05	0.948	0.922	0.934
Enrichr Pheno	OMIM EXP DIABETES MELLITUS TYPE 2	102	286	−0.1	−0.09	0.942	0.945	0.920
ESCAPE	ASCL1 19796622	1,349	88	0.2	0.01	0.745	0.706	0.667
ESCAPE	ATF3 19796622 DOWN	297	87	0.5	0.06	0.693	0.622	0.583
ESCAPE	CHIP CNOT3 19339689	1,153	104	1.5	0.40	0.714	0.648	0.664
ESCAPE	CHIP EED 16625203	633	129	0.1	0.00	0.853	0.825	0.801
ESCAPE	CHIP GCN5 20946988	223	114	2.2	0.48	0.719	0.575	0.673
ESCAPE	CHIP KLF4 18358816	1,303	109	1.1	0.30	0.726	0.672	0.670
ESCAPE	CHIP KLF4 18555785	1,908	105	1.0	0.28	0.744	0.686	0.709
ESCAPE	CHIP MYC 18555785	916	124	0.8	0.21	0.829	0.779	0.799
ESCAPE	CHIP NANOG 18347094	1,464	107	0.9	0.26	0.716	0.671	0.635
ESCAPE	CHIP NANOG 18358816	902	106	1.3	0.35	0.692	0.612	0.653
ESCAPE	CHIP PRDM14 21183938	1,448	101	1.2	0.29	0.715	0.678	0.657
ESCAPE	CHIP REST 18959480	1,752	99	0.5	0.14	0.715	0.652	0.653
ESCAPE	CHIP REST 21632747	1,554	80	0.2	−0.03	0.700	0.668	0.540
ESCAPE	EOMES 19796622	1,339	86	0.4	0.07	0.730	0.669	0.690
ESCAPE	EOMES 19796622 DOWN	1,357	79	0.3	0.01	0.711	0.660	0.635
ESCAPE	ESRRB 19136965	185	95	0.3	0.14	0.753	0.680	0.610
ESCAPE	ETV3 19796622	374	80	−0.1	−0.02	0.701	0.687	0.591
ESCAPE	FOXJ2 19796622	147	80	−0.3	−0.03	0.698	0.647	0.548
ESCAPE	GATA3 19796622 DOWN	1,608	83	0.4	0.08	0.708	0.662	0.615
ESCAPE	KLF4 18264089 DOWN	620	100	1.2	0.33	0.715	0.612	0.627
ESCAPE	KLF5 20875108 DOWN	209	131	0.0	0.06	0.821	0.778	0.699
ESCAPE	MESC H3K9ME3 19884255	1,925	102	0.3	0.07	0.744	0.717	0.684
ESCAPE	MYCN 19796622	361	73	0.6	0.06	0.707	0.640	0.536
ESCAPE	NANOG 16767105	419	89	0.6	0.15	0.706	0.638	0.652
ESCAPE	NR0B1 19530134 DOWN	147	107	1.8	0.36	0.727	0.622	0.669
ESCAPE	NR2F2 19796622 DOWN	823	78	0.5	0.05	0.717	0.696	0.625
ESCAPE	NR5A2 19796622	621	78	0.2	−0.01	0.715	0.647	0.498
ESCAPE	PANCT1 22327834 DOWN	278	89	0.6	0.15	0.768	0.734	0.618
ESCAPE	POU5F1 16767105	389	92	0.6	0.17	0.732	0.658	0.627
ESCAPE	PROTEIN NANOG 21589869	120	212	0.8	0.20	0.932	0.943	0.912
ESCAPE	PROTEIN POU5F1 22083510	187	196	0.2	0.05	0.915	0.856	0.890
ESCAPE	RAD21 21589869	505	98	1.6	0.33	0.718	0.662	0.642
ESCAPE	REST 21632747	141	97	4.6	0.43	0.702	0.549	0.673
ESCAPE	RHOX6 19796622	304	90	0.3	0.09	0.710	0.662	0.500
ESCAPE	SOX17 20123909 DOWN	267	109	0.2	0.06	0.699	0.675	0.567
ESCAPE	STAT3 19796622 DOWN	137	65	0.3	0.03	0.655	0.583	0.512
ESCAPE	T 19796622	696	89	0.2	0.07	0.719	0.666	0.592

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
ESCAPE	TCEA3 19796622	279	79	−0.3	−0.11	0.698	0.634	0.520
ESCAPE	TCEA3 19796622 DOWN	396	74	0.0	−0.07	0.696	0.648	0.589
ESCAPE	ZFP281 21915945 DOWN	157	172	1.6	0.44	0.896	0.798	0.835
GeneSigDB	12747878 TABLE2	152	126	2.9	0.44	0.761	0.716	0.732
GeneSigDB	12917485 TABLE6	451	134	2.1	0.45	0.797	0.701	0.709
GeneSigDB	12917485 TABLE9	384	143	2.7	0.51	0.783	0.713	0.753
GeneSigDB	15930337 TABLES4	152	122	5.3	0.69	0.737	0.627	0.715
GeneSigDB	16081686 SUPPTABLE4	640	174	0.9	0.30	0.849	0.800	0.817
GeneSigDB	16140871 SUPPTABLE7	154	107	0.1	0.08	0.836	0.815	0.742
GeneSigDB	16166618 SUPPTABLE4	1,430	98	0.5	0.16	0.752	0.694	0.713
GeneSigDB	16440291 SUPPTABLE1	195	115	0.1	0.00	0.854	0.829	0.779
GeneSigDB	16651414 SUPP2	1,606	112	0.9	0.25	0.731	0.696	0.688
GeneSigDB	16728981 SUPPTABLE2	113	111	6.4	0.69	0.658	0.563	0.632
GeneSigDB	16735486 TABLE1	126	101	0.9	0.28	0.789	0.733	0.746
GeneSigDB	16760443 SUPPTABLE2	148	161	3.6	0.43	0.781	0.693	0.750
GeneSigDB	16790086 TABLEW1	223	193	0.6	0.13	0.904	0.840	0.867
GeneSigDB	16872506 SUPPTABLE1	1,467	122	0.7	0.19	0.818	0.765	0.799
GeneSigDB	17555561 TABLE1	966	94	2.5	0.50	0.675	0.623	0.635
GeneSigDB	17597811 SUPPTABLE6	654	101	1.4	0.38	0.721	0.672	0.681
GeneSigDB	17671232 TABLES2A	170	110	0.2	0.12	0.732	0.673	0.652
GeneSigDB	17683608 TABLES2	590	106	1.1	0.30	0.752	0.689	0.713
GeneSigDB	17894856 SUPPLIST4	203	206	2.1	0.24	0.848	0.751	0.768
GeneSigDB	18062813 GENELIST	103	128	4.6	0.38	0.779	0.603	0.756
GeneSigDB	18394172 S3GENELIST	214	103	0.2	0.10	0.800	0.742	0.731
GeneSigDB	18394172 S4GENELIST	285	92	0.2	0.05	0.699	0.604	0.622
GeneSigDB	18535662 TABLES2A	871	107	0.6	0.21	0.770	0.698	0.720
GeneSigDB	18537972 TABLES3	649	108	2.7	0.55	0.692	0.599	0.597
GeneSigDB	18593951 TABLES2	214	102	5.2	0.53	0.687	0.622	0.637
GeneSigDB	18662380 S3 ESR1	269	91	1.0	0.26	0.725	0.639	0.622
GeneSigDB	18667080 TABLES2	743	129	1.3	0.37	0.770	0.683	0.739
GeneSigDB	18927307 TABLES2	236	113	5.6	0.67	0.711	0.552	0.680
GeneSigDB	18974375 TABLES2	165	103	0.1	0.09	0.847	0.772	0.794
GeneSigDB	19043454 TABLES2	166	121	−0.1	−0.02	0.873	0.860	0.820
GeneSigDB	19061838 TABLES14	199	118	3.3	0.37	0.762	0.706	0.698
GeneSigDB	19096012 TABLES2	427	128	1.1	0.35	0.768	0.696	0.729
GeneSigDB	19185848 TABLES4	206	104	0.2	0.09	0.789	0.725	0.746
GeneSigDB	19399471 SUPPTABLE2	101	109	3.6	0.55	0.643	0.494	0.629
GeneSigDB	19904269 ST1	205	104	1.1	0.33	0.695	0.640	0.622
GeneSigDB	20035825 TABLES7A	305	86	0.1	0.04	0.723	0.639	0.602
GeneSigDB	20215513 TABLES4	177	122	3.6	0.51	0.760	0.656	0.704
GeneSigDB	20421987 TABLES4	381	82	0.1	−0.01	0.727	0.657	0.612
GeneSigDB	20436685 ST5 2	214	106	3.8	0.46	0.748	0.680	0.637
GeneSigDB	20485376 TABLES4	149	1	0.0	0.00	0.998	0.994	0.995
GEO	AGING DN RAT HIPPOCAMPUS CA3 18 MONTHS VS 28 MONTHS GSE21681 AG	307	109	1.1	0.28	0.808	0.733	0.744
GEO	DIS PERT DN DILATED CARDIOMYOPATHY DOID 12930 HUMAN GSE42955 SA	295	92	0.4	0.18	0.717	0.639	0.556
GEO	DIS PERT DN ENDOMETRIOSIS DOID 289 HUMAN GSE6364 SAMPLE 947	379	104	0.6	0.15	0.788	0.734	0.752
GEO	DIS PERT DN MELANOMA DOID 1909 HUMAN GSE6887 SAMPLE 951	368	84	0.2	−0.03	0.713	0.668	0.672
GEO	DIS PERT DN MENTAL RETARDATION DOID 1059 HUMAN GSE6575 SAMPLE 1	414	79	0.3	0.07	0.678	0.658	0.626
GEO	DIS PERT UP ALS AMYOTROPHIC LATERAL SCLEROSIS C0002736 MOUSE	204	132	2.3	0.40	0.795	0.770	0.734
GEO	DIS PERT UP MORBID OBESITY DOID 11981 HUMAN GSE48964 SAMPLE 583	292	115	0.3	0.13	0.759	0.693	0.560

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
GEO	DIS PERT UP MULTIPLE SCLEROSIS DOID 2377 HUMAN GSE16461 SAMPLE	143	219	1.2	0.41	0.845	0.772	0.821
GEO	DRUG DN CISPLATIN DB00515 HUMAN GSE47856 SAMPLE 3150	162	95	0.3	0.09	0.695	0.700	0.585
GEO	DRUG DN CISPLATIN DB00515 HUMAN GSE47856 SAMPLE 3153	299	75	0.4	0.01	0.624	0.605	0.521
GEO	DRUG DN COLORECTAL ADENOCARCINOMA DB00482 HUMAN GSE11237 SAMPLE	293	107	0.0	0.05	0.765	0.702	0.736
GEO	DRUG DN TESTOSTERONE 6013 HUMAN GSE5106 SAMPLE 3206	325	152	1.4	0.25	0.840	0.800	0.804
GEO	DRUG PERT OLANZAPINE RATTUS NORVEGICUS GPL1355 GDS2608 CHDIR	141	121	2.4	0.38	0.788	0.710	0.736
GEO	DRUG UP DICLOFENAC DB00586 HUMAN GSE54255 SAMPLE 3053	350	123	3.0	0.41	0.694	0.661	0.673
GEO	DRUG UP ESTRADIOL 5757 HUMAN GSE26834 SAMPLE 3241	328	120	0.3	0.16	0.784	0.761	0.723
GEO	GENE DN ACVR1 KO MOUSE GSE46689 SAMPLE 2520	234	176	1.1	0.42	0.852	0.810	0.817
GEO	GENE DN GATA5 KO MOUSE GSE47425 SAMPLE 413	311	121	3.0	0.48	0.738	0.656	0.708
GEO	GENE DN NR1I3 KO MOUSE GSE40120 SAMPLE 3066	267	112	0.2	0.10	0.808	0.732	0.745
GEO	GENE DN PAFAH1B1 KD MOUSE GSE35366 SAMPLE 1618	236	75	0.3	0.01	0.642	0.610	0.558
GEO	GENE DN SON KD HUMAN GSE26888 SAMPLE 62	273	119	1.7	0.35	0.811	0.743	0.800
GEO	GENE DN TRP53 KO MOUSE GSE40545 SAMPLE 1829	237	104	0.0	0.03	0.818	0.793	0.734
GEO	GENE UP CCAR2 KD HUMAN GSE54707 SAMPLE 1072	175	84	4.5	0.42	0.659	0.618	0.601
GEO	GENE UP CDH1 KO MOUSE GSE48131 SAMPLE 2982	297	77	0.2	0.04	0.756	0.726	0.679
GEO	GENE UP DMRTA2 OE MOUSE GSE25179 SAMPLE 2204	175	110	0.2	0.06	0.748	0.689	0.635
GEO	GENE UP HNF4A DEPLETION HUMAN GSE29084 SAMPLE 118	351	63	0.0	−0.10	0.685	0.636	0.575
GEO	GENE UP ITK KNOCKOUT MOUSE GSE12465 SAMPLE 1995	257	116	2.4	0.39	0.779	0.757	0.705
GEO	GENE UP MIR26B OE HUMAN GSE12091 SAMPLE 2978	303	90	0.6	0.19	0.661	0.597	0.497
GEO	GENE UP STK19 OE HUMAN GSE36036 SAMPLE 1570	335	95	0.2	0.12	0.701	0.603	0.517
GEO	GENE UP TWIST1 OE MOUSE GSE50002 SAMPLE 1075	330	91	1.1	0.29	0.685	0.647	0.549

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
GEO	LIGAND DN ESTROGEN MOUSE ALPHA NULL AORTA GDS3064 LIGAND 26	396	71	0.1	−0.06	0.809	0.813	0.749
GEO	LIGAND UP 17BETA ESTRADIOL MOUSE PROSTATE GLAND GSE36630 LIGAND	311	75	0.1	0.03	0.707	0.671	0.529
GEO	LIGAND UP FIBROBLAST GROWTH FACTOR 2 FGF2 HUMAN EMBRYONIC FIB	333	130	2.2	0.37	0.801	0.779	0.756
GEO	MCF7 DN MAP2K6 SIRNA 48 HRS GSE53668 MCF7 276	336	145	1.3	0.39	0.849	0.799	0.826
GEO	MCF7 UP BORTEZOMIB VELCADE GSE30931 MCF7 16	360	119	0.7	0.18	0.779	0.716	0.746
GEO	MICROBE UP STREPTO-COCCUS PNEUMONIAE D39 HUMAN PHARYNGEAL EPITHE	443	158	0.9	0.24	0.845	0.789	0.805
GEO	TF LOF CDX2 20696899 CACO2 LOF HUMAN GPL570 GSE22572	1,352	91	1.8	0.40	0.653	0.590	0.617
GEO	TF LOF GRHL3 16949565 SKIN LOF MOUSE GPL1261 GDS2629 DOWN	168	83	5.0	0.51	0.654	0.584	0.603
GEO	VIRUS DN ICSARA DELTAORF6 60HOUR GSE33267	286	75	4.6	0.56	0.618	0.551	0.527
GEO	VIRUS UP HCV JFH 1 18HOUR GSE20948	294	101	1.6	0.31	0.754	0.647	0.650
GEO	VIRUS UP SARS COV 0HOUR GSE47960	290	90	3.4	0.53	0.672	0.565	0.630
GO (test)	GO:0000184	118	385	0.0	0.06	0.990	0.996	0.991
GO (test)	GO:0000398	239	138	0.1	0.00	0.989	0.993	0.985
GO (test)	GO:0003676	255	79	6.9	0.42	0.926	0.885	0.943
GO (test)	GO:0003924	280	180	0.0	−0.03	0.990	0.987	0.994
GO (test)	GO:0004222	116	77	0.2	−0.04	0.990	0.998	0.995
GO (test)	GO:0005096	283	91	0.1	0.07	0.979	0.977	0.973
GO (test)	GO:0005739	1,291	87	0.2	0.08	0.930	0.863	0.916
GO (test)	GO:0005802	131	91	0.4	0.16	0.913	0.943	0.968
GO (test)	GO:0005814	118	82	0.5	0.14	0.918	0.971	0.977
GO (test)	GO:0005925	392	201	0.6	0.14	0.962	0.911	0.973
GO (test)	GO:0006413	133	353	0.0	0.08	0.987	0.994	0.982
GO (test)	GO:0006511	158	138	0.1	0.13	0.979	0.985	0.975
GO (test)	GO:0006898	189	125	−0.4	0.00	0.919	0.956	0.919
GO (test)	GO:0006914	110	133	0.5	0.12	0.918	0.969	0.948
GO (test)	GO:0007156	156	125	0.0	0.09	0.995	0.998	0.992
GO (test)	GO:0007186	862	273	0.5	0.03	0.990	0.973	0.980
GO (test)	GO:0007204	129	193	0.1	0.07	0.968	0.984	0.971
GO (test)	GO:0007568	165	205	0.0	−0.01	0.908	0.901	0.925
GO (test)	GO:0008022	177	227	1.7	0.21	0.923	0.906	0.978
GO (test)	GO:0010629	136	223	1.5	0.22	0.919	0.910	0.962
GO (test)	GO:0016323	183	118	−0.1	−0.10	0.948	0.919	0.980
GO (test)	GO:0016567	430	127	0.0	0.02	0.978	0.966	0.984
GO (test)	GO:0016887	153	150	0.0	0.00	0.979	0.965	0.988
GO (test)	GO:0017124	122	134	0.3	0.07	0.930	0.966	0.990
GO (test)	GO:0030168	109	222	−0.3	−0.14	0.978	0.969	0.984
GO (test)	GO:0030198	195	133	0.0	0.11	0.975	0.950	0.988
GO (test)	GO:0031625	294	209	0.5	0.10	0.938	0.939	0.960
GO (test)	GO:0031901	127	107	0.5	0.16	0.917	0.981	0.964
GO (test)	GO:0033209	127	190	0.2	0.06	0.989	0.972	0.973
GO (test)	GO:0034220	206	87	0.0	0.00	0.990	0.984	0.982
GO (test)	GO:0043025	313	146	0.2	0.10	0.891	0.846	0.942
GO (test)	GO:0043066	461	187	0.8	0.15	0.944	0.848	0.957
GO (test)	GO:0046777	166	406	0.1	0.03	0.988	0.984	0.997
GO (test)	GO:0046854	103	159	0.3	0.06	0.979	0.972	0.969

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
GO (test)	GO:0048471	628	135	1.0	0.27	0.905	0.827	0.964
GO (test)	GO:0050900	205	128	0.0	−0.14	0.906	0.959	0.945
GO (test)	GO:0051015	145	131	0.0	0.06	0.944	0.971	0.979
GO (test)	GO:0051091	103	247	−0.3	−0.15	0.989	0.932	0.989
GO (test)	GO:0070374	180	201	0.6	0.13	0.942	0.939	0.949
GO (test)	GO:0072562	190	123	0.0	−0.14	0.960	0.945	0.990
LINCS DN	CPC001 HCC515 24H RS 17053 HYDROCHLORIDE 10 0	200	114	3.5	0.54	0.726	0.684	0.642
LINCS DN	CPC003 VCAP 24H SERICETIN DIMETHYL ETHER 10 0	116	102	1.4	0.22	0.693	0.563	0.670
LINCS DN	CPC005 A375 24H TRICHOSTATIN A 10 0	159	108	5.2	0.44	0.728	0.633	0.684
LINCS DN	CPC006 A375 24H IMD 0354 10 0	169	114	4.7	0.54	0.722	0.658	0.672
LINCS DN	CPC006 A549 24H TRICHOSTATIN A 10 0	115	101	0.1	0.05	0.738	0.646	0.660
LINCS DN	CPC006 HCC515 24H EMETINE DIHYDROCHLORIDE HYDRATE 74	223	115	4.2	0.62	0.707	0.544	0.656
LINCS DN	CPC006 LOVO 6H HDAC6 INHIBITOR ISOX 10 0	114	123	5.0	0.52	0.714	0.540	0.699
LINCS DN	CPC006 NCIH508 6H MINOXIDIL 10 0	118	109	3.9	0.47	0.728	0.621	0.706
LINCS DN	CPC006 PC3 24H MANUMYCIN A 10 0	228	110	4.1	0.51	0.733	0.724	0.635
LINCS DN	CPC006 THP1 6H BRD K92301463 10 0	122	125	6.6	0.58	0.682	0.641	0.688
LINCS DN	CPC006 VCAP 24H BI 2536 10 0	141	111	3.7	0.59	0.723	0.595	0.655
LINCS DN	CPC008 A375 24H 2 CHLORO 7 METHOXYPHENOTHIAZINE 10 0	158	103	4.9	0.56	0.681	0.631	0.655
LINCS DN	CPC008 VCAP 6H TRICHOSTATIN A 10 0	159	112	2.2	0.44	0.719	0.605	0.645
LINCS DN	CPC009 PC3 24H BRD K83670234 10 0	104	113	5.5	0.42	0.689	0.636	0.666
LINCS DN	CPC010 A375 6H WORTMANNIN 10 0	126	121	4.6	0.56	0.706	0.647	0.696
LINCS DN	CPC011 A549 6H TRICHOSTATIN A 10 0	129	112	3.2	0.37	0.654	0.480	0.647
LINCS DN	CPC011 MCF7 24H TOPOTECAN HCL 10 0	126	96	6.1	0.61	0.688	0.605	0.636
LINCS DN	CPC011 PC3 24H IDARUBICIN HCL 10 0	173	100	0.8	0.25	0.729	0.538	0.651
LINCS DN	CPC012 A549 24H TRICHOSTATIN A 10 0	130	117	2.6	0.32	0.743	0.645	0.700
LINCS DN	CPC012 ASC 24H K784 3187 10 0	101	113	2.8	0.38	0.697	0.642	0.646
LINCS DN	CPC012 HT29 6H BRD K87909389 10 0	138	123	6.0	0.58	0.737	0.641	0.695
LINCS DN	CPC013 VCAP 24H NP 007374 10 0	106	111	6.9	0.51	0.706	0.563	0.714
LINCS DN	CPC014 MCF7 24H BRD A20697603 10 0	150	135	4.4	0.59	0.719	0.610	0.671
LINCS DN	CPC018 A549 24H NSC 3852 10 0	121	111	4.5	0.56	0.710	0.607	0.686
LINCS DN	CPC018 HT29 6H MEK1 2 INHIBITOR 10 0	158	103	3.6	0.41	0.713	0.617	0.663
LINCS DN	CPD001 MCF7 24H IFENPRODIL TARTRATE 10 0	123	122	4.9	0.52	0.754	0.667	0.726
LINCS DN	CPD002 MCF7 24H GELDANAMYCIN 10 0	129	96	4.0	0.47	0.685	0.558	0.685
LINCS DN	LJP005 A375 24H CP 724714 10	104	217	3.9	0.33	0.858	0.786	0.848

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
LINCS DN	LJP005 A375 24H WITHAFERIN A 0 04	159	133	1.1	0.26	0.815	0.788	0.735
LINCS DN	LJP005 HA1E 24H NVP AUY922 0 12	136	145	−0.1	−0.02	0.838	0.770	0.787
LINCS DN	LJP005 HEPG2 24H GELDANAMYCIN 3 33	151	118	0.3	0.17	0.760	0.667	0.685
LINCS DN	LJP005 HEPG2 24H NVP BEZ235 0 37	129	156	0.4	0.17	0.903	0.812	0.864
LINCS DN	LJP005 HS578T 24H PD 0325901 0 37	120	157	0.1	0.07	0.857	0.869	0.801
LINCS DN	LJP005 PC3 24H MITOXANTRONE 10	111	162	4.2	0.43	0.803	0.584	0.786
LINCS DN	LJP006 A375 24H PF 562271 10	100	145	2.9	0.28	0.792	0.760	0.768
LINCS DN	LJP006 A549 24H A443654 0 37	155	110	1.1	0.31	0.773	0.707	0.719
LINCS DN	LJP006 HS578T 24H RADICICOL 10	122	122	−0.1	0.04	0.804	0.742	0.774
LINCS DN	LJP007 A375 24H NVP BGT226 0 37	161	116	0.9	0.27	0.804	0.699	0.736
LINCS DN	LJP009 MCF7 24H ON 01910 0 37	109	141	1.2	0.28	0.793	0.637	0.754
LINCS DN	LJP009 PC3 24H CGP 60474 0 37	124	140	2.9	0.44	0.778	0.699	0.743
MSigDB	ACEVEDO FGFR1 TARGETS IN PROSTATE CANCER MODEL	597	89	0.3	0.06	0.732	0.670	0.655
MSigDB	ACEVEDO LIVER CANCER	1,513	86	0.7	0.15	0.726	0.609	0.701
MSigDB	ACEVEDO LIVER CANCER WITH H3K27ME3	519	82	0.5	0.08	0.628	0.642	0.506
MSigDB	ACEVEDO LIVER CANCER WITH H3K9ME3	257	77	0.1	0.00	0.581	0.564	0.479
MSigDB	AGUIRRE PANCREATIC CANCER COPY NUMBER	534	113	1.4	0.35	0.730	0.657	0.667
MSigDB	BERTUCCI MEDULLARY VS DUCTAL BREAST CANCER	375	102	1.9	0.37	0.694	0.665	0.654
MSigDB	BOYAULT LIVER CANCER SUBCLASS G3	239	125	1.5	0.38	0.805	0.716	0.754
MSigDB	CAMPS COLON CANCER COPY NUMBER	165	107	0.1	0.00	0.690	0.654	0.596
MSigDB	CHARAFE BREAST CANCER LUMINAL VS BASAL	834	94	1.0	0.26	0.724	0.652	0.662
MSigDB	CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL	910	87	0.7	0.16	0.726	0.667	0.660
MSigDB	CHIANG LIVER CANCER SUBCLASS CTNNB1	346	88	0.2	0.04	0.766	0.706	0.701
MSigDB	CHIANG LIVER CANCER SUBCLASS PROLIFERATION	356	93	0.1	0.00	0.807	0.776	0.725
MSigDB	CHIANG LIVER CANCER SUBCLASS UNANNOTATED	278	122	1.5	0.40	0.774	0.710	0.708
MSigDB	DELYS THYROID CANCER	675	107	0.3	0.10	0.810	0.779	0.761
MSigDB	GINESTIER BREAST CANCER 20Q13 AMPLIFICATION	298	92	0.4	0.05	0.709	0.642	0.580
MSigDB	GINESTIER BREAST CANCER ZNF217 AMPLIFIED	413	84	0.6	0.09	0.708	0.629	0.671
MSigDB	GRADE COLON AND RECTAL CANCER	386	128	1.4	0.37	0.762	0.699	0.642
MSigDB	GRUETZMANN PANCREATIC CANCER	561	139	0.9	0.30	0.803	0.740	0.744
MSigDB	HOSHIDA LIVER CANCER LATE RECURRENCE	131	113	5.0	0.44	0.738	0.590	0.715
MSigDB	HOSHIDA LIVER CANCER SURVIVAL	186	116	0.1	0.09	0.767	0.717	0.682

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
MSigDB	HUMMERICH SKIN CANCER PROGRESSION	188	164	0.4	0.14	0.870	0.759	0.795
MSigDB	LAIHO COLORECTAL CANCER SERRATED	198	132	4.4	0.53	0.737	0.669	0.691
MSigDB	LEE LIVER CANCER ACOX1	129	89	0.0	0.07	0.801	0.698	0.733
MSigDB	LEE LIVER CANCER CIPROFIBRATE	126	97	−0.1	−0.03	0.830	0.759	0.759
MSigDB	LEE LIVER CANCER DENA	134	109	0.1	0.19	0.845	0.779	0.785
MSigDB	LEE LIVER CANCER E2F1	126	99	0.0	0.07	0.813	0.684	0.783
MSigDB	LEE LIVER CANCER MYC E2F1	120	105	−0.3	−0.03	0.833	0.783	0.775
MSigDB	LEE LIVER CANCER MYC TGFA	126	98	−0.3	−0.10	0.815	0.748	0.781
MSigDB	LEE LIVER CANCER SURVIVAL	360	104	0.9	0.22	0.759	0.727	0.721
MSigDB	LINDGREN BLADDER CANCER CLUSTER 1	499	113	2.1	0.40	0.748	0.652	0.686
MSigDB	LINDGREN BLADDER CANCER CLUSTER 3	558	113	1.5	0.36	0.756	0.652	0.686
MSigDB	LIU PROSTATE CANCER	577	102	0.2	0.04	0.756	0.672	0.672
MSigDB	OSMAN BLADDER CANCER	804	107	1.3	0.31	0.739	0.686	0.725
MSigDB	POOLA INVASIVE BREAST CANCER	422	115	0.2	0.13	0.801	0.726	0.722
MSigDB	ROESSLER LIVER CANCER METASTASIS	160	119	5.9	0.52	0.699	0.509	0.676
MSigDB	SCHUETZ BREAST CANCER DUCTAL INVASIVE	435	114	0.3	0.17	0.792	0.759	0.730
MSigDB	SMID BREAST CANCER BASAL	1,349	98	0.5	0.15	0.746	0.692	0.661
MSigDB	SMID BREAST CANCER LUMINAL B	736	102	0.2	0.02	0.784	0.710	0.665
MSigDB	SMID BREAST CANCER RELAPSE IN BONE	412	107	0.3	0.14	0.759	0.736	0.681
MSigDB	SOTIRIOU BREAST CANCER GRADE 1 VS 3	203	153	−0.1	−0.04	0.871	0.844	0.794
MSigDB	STEARMAN LUNG CANCER EARLY VS LATE	186	130	4.6	0.47	0.743	0.597	0.688
MSigDB	SWEET LUNG CANCER KRAS	921	110	0.6	0.17	0.785	0.727	0.736
MSigDB	VANTVEER BREAST CANCER ESR1	407	95	1.3	0.31	0.699	0.587	0.653
MSigDB	VANTVEER BREAST CANCER METASTASIS	177	96	0.3	0.09	0.741	0.640	0.640
MSigDB	VECCHI GASTRIC CANCER ADVANCED VS EARLY	138	62	−0.2	−0.08	0.697	0.708	0.701
MSigDB	VECCHI GASTRIC CANCER EARLY	795	90	0.4	0.10	0.705	0.658	0.656
MSigDB	WALLACE PROSTATE CANCER RACE	387	100	0.1	0.08	0.759	0.746	0.678
MSigDB	WAMUNYOKOLI OVARIAN CANCER GRADES 1 2	204	77	0.0	−0.03	0.728	0.697	0.649
MSigDB	WAMUNYOKOLI OVARIAN CANCER LMP	464	79	0.6	0.20	0.680	0.633	0.661
MSigDB	WANG ESOPHAGUS CANCER VS NORMAL	222	114	0.6	0.20	0.788	0.701	0.688
MSigDB	WATANABE RECTAL CANCER RADIOTHERAPY RESPONSIVE	200	160	2.6	0.57	0.771	0.628	0.765
MSigDB	WOO LIVER CANCER RECURRENCE	185	120	0.0	0.02	0.890	0.836	0.823
MSigDB	ZHANG BREAST CANCER PROGENITORS	566	106	1.0	0.27	0.750	0.679	0.697
Pathcom	2 2 AMINO 3 METHOXYPHENYL 4H 1 BENZOPYRAN 4 ONE INHIBITS	233	261	−0.1	−0.06	0.941	0.928	0.922

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Pathcom	ABC FAMILY PROTEINS MEDIATED TRANSPORT ATP SENSITIVE POTAS	131	79	0.1	0.07	0.996	0.982	0.988
Pathcom	CTTTGA V LEF1 Q2 CTTTGT V LEF1 Q2	322	128	0.2	0.09	0.799	0.719	0.701
Pathcom	CTTTGT V LEF1 Q2	1,493	104	0.8	0.23	0.714	0.662	0.651
Pathcom	GGGTGGRR V PAX4 03	900	115	1.2	0.30	0.735	0.686	0.690
Pathcom	RCGCANGCGY V NRF1 Q6 V NRF1 Q6	203	102	3.0	0.42	0.711	0.639	0.678
Pathcom	RGAANNTTC V HSF1 01	209	107	4.9	0.57	0.717	0.583	0.624
Pathcom	SGCGSSAAA V E2F1DP2 01 V E2F1DP1RB 01	120	112	0.4	0.14	0.815	0.786	0.788
Pathcom	TGACCTTG V SF1 Q6 V SF1 Q6	136	96	−0.2	−0.08	0.694	0.583	0.638
Pathcom	TGACCTY V ERR1 Q2	811	104	0.9	0.25	0.701	0.635	0.653
Pathcom	TGACCTY V ERR1 Q2 V ERR1 Q2	170	102	0.1	0.00	0.713	0.623	0.644
Pathcom	TGCCAAR V NF1 Q6	544	102	1.1	0.30	0.713	0.606	0.590
Pathcom	TGF BETA RECEPTOR	185	363	0.2	0.09	0.964	0.956	0.921
Pathcom	V AP4 Q6 01	134	128	2.8	0.41	0.750	0.647	0.692
Pathcom	V AREB6 01	100	121	0.6	0.24	0.680	0.589	0.615
Pathcom	V CEBPB 01	179	126	3.5	0.45	0.730	0.663	0.649
Pathcom	V CREB Q3	140	127	4.6	0.56	0.702	0.610	0.695
Pathcom	V DBP Q6	248	106	−0.1	−0.04	0.747	0.675	0.628
Pathcom	V EN1 01	104	151	0.4	0.15	0.764	0.677	0.743
Pathcom	V ER Q6 01	225	115	1.1	0.32	0.726	0.597	0.642
Pathcom	V FOXD3 01	188	123	1.7	0.43	0.754	0.759	0.684
Pathcom	V GATA2 01	100	153	6.0	0.44	0.697	0.680	0.694
Pathcom	V HIF1 Q3 V HIF1 Q5	126	117	0.2	0.07	0.732	0.665	0.692
Pathcom	V HNF6 Q6	228	118	0.4	0.16	0.749	0.655	0.631
Pathcom	V HOXA4 Q2	239	135	0.6	0.24	0.761	0.698	0.635
Pathcom	V IRF1 01	137	119	2.1	0.29	0.738	0.612	0.706
Pathcom	V IRF2 01	120	109	3.2	0.45	0.721	0.648	0.714
Pathcom	V LBP1 Q6	199	92	2.7	0.44	0.719	0.638	0.611
Pathcom	V LMO2COM 01	220	105	3.2	0.44	0.750	0.648	0.617
Pathcom	V MYB Q6	107	116	7.8	0.57	0.703	0.686	0.657
Pathcom	V MYCMAX 02	106	117	1.4	0.24	0.718	0.649	0.699
Pathcom	V PAX4 02	143	86	−0.4	−0.09	0.719	0.710	0.573
Pathcom	V PTF1BETA Q6	221	119	1.6	0.29	0.732	0.623	0.655
Pathcom	V SOX9 B1	113	132	0.0	0.13	0.771	0.645	0.720
Pathcom	V SP3 Q3	234	114	2.1	0.38	0.718	0.678	0.620
Pathcom	V SRY 02	194	133	−0.2	−0.10	0.780	0.672	0.662
Pathcom	V TAL1BETAITF2 01	217	91	0.2	0.07	0.732	0.646	0.584
Pathcom	V TCF11 01	242	118	1.4	0.36	0.711	0.635	0.609
Pathcom	V TFIIA Q6	225	119	1.3	0.31	0.733	0.631	0.632
Pathcom	WNT	101	302	−0.4	−0.14	0.970	0.976	0.946
Reactome	HSA 112314	154	127	0.0	0.02	0.980	0.981	0.961
Reactome	HSA 112412	265	204	0.1	0.06	0.954	0.983	0.923
Reactome	HSA 1428517	177	113	0.0	−0.02	0.976	0.998	0.960
Reactome	HSA 1500931	141	156	1.5	0.26	0.984	0.991	0.980
Reactome	HSA 162906	260	155	0.3	0.09	0.922	0.979	0.906
Reactome	HSA 1630316	125	70	−0.3	−0.04	0.986	0.999	0.975
Reactome	HSA 166520	510	207	0.1	0.02	0.968	0.964	0.907
Reactome	HSA 168179	106	354	0.1	0.02	0.944	0.992	0.946
Reactome	HSA 1852241	418	112	0.3	0.12	0.960	0.962	0.938
Reactome	HSA 194840	148	120	−0.1	−0.04	0.995	0.997	0.981
Reactome	HSA 195258	314	151	0.0	0.01	0.979	0.967	0.957
Reactome	HSA 198203	135	260	0.2	0.03	0.981	0.955	0.963
Reactome	HSA 202424	156	165	0.0	−0.02	0.988	0.984	0.986
Reactome	HSA 211945	114	109	0.0	0.01	0.975	0.994	0.985
Reactome	HSA 2172127	424	220	−0.1	−0.01	0.895	0.972	0.897
Reactome	HSA 2428924	319	212	0.1	0.05	0.962	0.981	0.893
Reactome	HSA 2555396	214	144	0.1	0.05	0.949	0.991	0.927
Reactome	HSA 2871796	329	181	0.1	0.05	0.883	0.973	0.884
Reactome	HSA 373080	101	132	0.1	0.05	0.996	0.999	0.988
Reactome	HSA 375165	302	189	0.1	0.05	0.962	0.977	0.898
Reactome	HSA 397014	204	126	0.5	0.20	0.977	0.991	0.972

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
Reactome	HSA 418594	280	113	0.3	0.04	0.970	0.993	0.928
Reactome	HSA 446652	111	258	0.0	0.05	0.967	0.998	0.945
Reactome	HSA 448424	328	231	0.1	0.05	0.960	0.981	0.898
Reactome	HSA 453279	165	252	0.0	−0.01	0.951	0.993	0.903
Reactome	HSA 5368287	103	78	0.0	0.03	0.985	0.996	0.943
Reactome	HSA 5578749	105	135	0.0	0.02	0.982	0.996	0.994
Reactome	HSA 5683057	317	212	0.2	0.09	0.956	0.978	0.849
Reactome	HSA 5684996	271	205	0.1	0.07	0.961	0.983	0.889
Reactome	HSA 6803157	107	26	3.5	0.21	0.952	0.981	0.939
Reactome	HSA 6811434	104	137	0.0	0.03	0.986	0.995	0.982
Reactome	HSA 68886	347	138	0.0	0.03	0.931	0.988	0.940
Reactome	HSA 69002	102	207	−0.3	−0.14	0.945	0.999	0.915
Reactome	HSA 73864	133	125	0.1	0.04	0.983	0.996	0.930
Reactome	HSA 74752	349	190	0.1	0.03	0.952	0.979	0.917
Reactome	HSA 8853659	290	205	0.1	0.06	0.961	0.978	0.904
Reactome	HSA 8953854	804	142	0.1	0.02	0.926	0.969	0.887
Reactome	HSA 912526	276	205	0.2	0.07	0.951	0.980	0.887
Reactome	HSA 975957	126	372	0.0	0.01	0.995	0.995	0.985
Reactome	HSA 983169	445	133	0.0	0.02	0.891	0.975	0.912
TargetScan	AAAGACA MIR 511	169	125	2.6	0.38	0.729	0.621	0.641
TargetScan	AAAGGAT MIR 501	106	110	0.9	0.25	0.755	0.645	0.712
TargetScan	AAGCACA MIR 218	337	120	0.5	0.25	0.754	0.686	0.690
TargetScan	AAGCACT MIR 520F	196	119	1.1	0.30	0.766	0.623	0.677
TargetScan	ACACTGG MIR 199A MIR 199B	136	148	2.9	0.30	0.794	0.706	0.741
TargetScan	ACATTCC MIR 1 MIR 206	247	125	3.0	0.44	0.762	0.688	0.705
TargetScan	ACCATTT MIR 522	145	133	1.4	0.27	0.775	0.675	0.725
TargetScan	ACTGAAA MIR 30A 3P MIR 30E 3P	164	146	0.6	0.18	0.793	0.679	0.713
TargetScan	ACTGCCT MIR 34B	190	112	0.5	0.21	0.755	0.693	0.687
TargetScan	ACTGTGA MIR 27A MIR 27B	388	111	2.5	0.46	0.757	0.691	0.626
TargetScan	AGCACTT MIR 93 MIR 302A MIR 302B MIR 302C MIR 302D M	282	121	0.3	0.09	0.767	0.723	0.734
TargetScan	AGCATTA MIR 155	116	134	3.0	0.26	0.798	0.727	0.752
TargetScan	ATAAGCT MIR 21	103	135	2.3	0.39	0.798	0.827	0.791
TargetScan	ATATGCA MIR 448	190	120	0.6	0.25	0.727	0.652	0.635
TargetScan	ATGAAGG MIR 205	134	113	0.4	0.20	0.755	0.679	0.707
TargetScan	ATGTAGC MIR 221 MIR 222	114	127	0.3	0.18	0.743	0.657	0.751
TargetScan	CACTGCC MIR 34A MIR 34C MIR 449	252	100	1.1	0.28	0.744	0.672	0.599
TargetScan	CAGCACT MIR 512 3P	134	112	−0.1	−0.07	0.744	0.726	0.642
TargetScan	CATGTAA MIR 496	152	115	0.1	0.06	0.740	0.667	0.682
TargetScan	CCTGCTG MIR 214	194	103	−0.2	0.01	0.720	0.628	0.665
TargetScan	CCTGTGA MIR 513	109	112	5.5	0.64	0.702	0.690	0.687
TargetScan	CTCTGGA MIR 520A MIR 525	133	122	1.9	0.27	0.715	0.664	0.732
TargetScan	CTTTGCA MIR 527	202	130	2.5	0.37	0.776	0.703	0.733
TargetScan	GACTGTT MIR 212 MIR 132	133	101	1.9	0.34	0.758	0.657	0.728
TargetScan	GAGCCAG MIR 149	122	113	4.4	0.50	0.734	0.583	0.683
TargetScan	GAGCTGG MIR 337	133	98	0.2	0.05	0.700	0.579	0.676
TargetScan	GCATTTG MIR 105	150	114	5.5	0.46	0.724	0.557	0.714
TargetScan	GGCACTT MIR 519E	108	135	0.4	0.15	0.789	0.718	0.747
TargetScan	GGGACCA MIR 133A MIR 133B	176	121	1.9	0.33	0.676	0.699	0.599
TargetScan	GGGCATT MIR 365	101	117	−0.1	0.00	0.777	0.698	0.741
TargetScan	GTACTGT MIR 101	219	138	1.6	0.32	0.792	0.687	0.729
TargetScan	GTGCAAA MIR 507	109	149	0.4	0.20	0.781	0.642	0.742
TargetScan	TCTCTCC MIR 185	102	121	0.1	−0.01	0.742	0.565	0.696
TargetScan	TGCACTT MIR 519C MIR 519B MIR 519A	377	123	1.2	0.33	0.759	0.715	0.698
TargetScan	TGCCTTA MIR 124A	474	95	0.7	0.21	0.761	0.695	0.681
TargetScan	TGCTGCT MIR 15A MIR 16 MIR 15B MIR 195 MIR 424 MIR 4	518	118	0.8	0.27	0.735	0.669	0.662
TargetScan	TGGTGCT MIR 29A MIR 29B MIR 29C	438	109	0.5	0.19	0.745	0.683	0.661

Continued on next page →

Table A.1 – continued from previous page

Collection	Gene Set	Size	Median Degree	ΔMed	Δ200+	MAPR AUC	DRaWR AUC	LASSO AUC
TargetScan	TGTTTAC MIR 30A 5P MIR 30C MIR 30D MIR 30B MIR 30E 5	494	105	0.3	0.15	0.752	0.675	0.667
TargetScan	TTTGCAG MIR 518A 2	179	123	0.2	0.11	0.782	0.696	0.713
TargetScan	TTTTGAG MIR 373	195	126	4.1	0.48	0.765	0.658	0.722

End of table.

APPENDIX B

COMPARISONS OF SIMILARITY RANK AND MUTATION RATES FOR GENE FAMILIES IN BTNR

Two tables are presented here for each of the following three gene families found to be highly ranked for BTNR: claudins, kallikreins and collagen type N alpha chains.

The first table in each pair presents the gene name and brief description, followed and sorted by the GeneSet MAPR similarity rank (out of 23,782). It also shows whether a gene was part of the original 323 genes in BTNR as well as the log fold change in DE and p-value measured as part of the phenotype created during the BEAUTY study. Finally, for of the other 9 cancer gene sets tested for comparison, the table shows how many sets in which the gene appeared and the median MAPR similarity rank.

The second table in each pair again shows the gene name, MAPR similarity rank, and membership in BTNR. It also shows SNV and CNV mutation rates measured during the BEAUTY study for participants with triple-negative breast cancer, separated by non-responders (nR) and pathological complete response (pCR). Additionally, it shows mutation rates from TCGA across breast cancer cases (BC) and all cases, as well as whether the gene is identified by COSMIC as a Tier 1 or Tier 2 cancer gene.

Table B.1: Summary of Ranked Claudins for BTNR using GeneSet MAPR, Part 1 of 2

Gene Name	Description	MAPR rank	BTNR in set	BTNR logFC	BTNR p-val	Others in set	Others med rnk
CLDN5	claudin 5	2		0.04	0.901	1	2,998
CLDN22	claudin 22	4					6,280
CLDN2	claudin 2	9	y	0.99	0.023	1	1,046
CLDN8	claudin 8	17	y	1.11	0.029	2	8,759
CLDN17	claudin 17	22					7,783
CLDN23	claudin 23	23		−0.31	0.348		9,401
CLDN19	claudin 19	32		0.71	0.189		7,434
CLDN15	claudin 15	46		0.20	0.346		4,917
CLDN4	claudin 4	69		0.77	0.005		6,374
CLDN11	claudin 11	71		0.63	0.091		7,601
CLDN3	claudin 3	75		0.68	0.100		5,029
CLDN18	claudin 18	78	y	0.85	0.012		4,933
CLDN10	claudin 10	117	y	1.28	0.054	2	2,268
CLDN14	claudin 14	119		−0.16	0.682		6,328
CLDN16	claudin 16	196		0.57	0.233		6,160
CLDN7	claudin 7	226		0.22	0.512		2,610
CLDN12	claudin 12	240		0.33	0.033	1	4,654
CLDN6	claudin 6	284	y	1.58	0.002	1	7,542
CLDN9	claudin 9	373		−0.88	0.038		5,199
CLDN1	claudin 1	386	y	0.88	0.079	2	4,059
CLDN20	claudin 20	622		0.11	0.773	1	9,639
CLDN25	claudin 25	688					10,451
CLDN24	claudin 24	3,674					17,373
CLDND2	claudin domain containing 2	9,722		−0.57	0.041		17,026
CLDN34	claudin 34	10,532					17,939
CLDND1	claudin domain containing 1	11,639		0.05	0.729		16,056

Table B.2: Summary of Ranked Claudins for BTNR using GeneSet MAPR, Part 2 of 2

Gene Name	MAPR rank	BTNR in set	SNV nR	SNV pCR	CNV nR	CNV pCR	TCGA, BC	TCGA, all	COSMIC Tier
CLDN5	2		0.011	0.000				0.005	
CLDN22	4							0.005	
CLDN2	9	y						0.010	
CLDN8	17	y						0.011	
CLDN17	22							0.011	
CLDN23	23				0.034	0.023		0.006	
CLDN19	32							0.008	
CLDN15	46							0.006	
CLDN4	69							0.010	
CLDN11	71				0.045	0.023		0.008	
CLDN3	75							0.003	
CLDN18	78	y						0.014	
CLDN10	117	y						0.012	
CLDN14	119							0.006	
CLDN16	196								
CLDN7	226							0.006	
CLDN12	240							0.010	
CLDN6	284	y						0.009	
CLDN9	373								
CLDN1	386	y						0.008	
CLDN20	622							0.006	
CLDN25	688								
CLDN24	3,674								
CLDND2	9,722								
CLDN34	10,532								
CLDND1	11,639								

Table B.3: Summary of Ranked Kallikreins for BTNR using GeneSet MAPR, Part 1 of 2

Gene Name	Description	MAPR rank	BTNR in set	BTNR logFC	BTNR p-val	Others in set	Others med rnk
KLK13	kallikrein related peptidase 13	20	y	−1.45	0.013	1	7,699
KLK8	kallikrein related peptidase 8	28	y	−2.00	0.002	3	7,367
KLK12	kallikrein related peptidase 12	30					10,077
KLK7	kallikrein related peptidase 7	35	y	−2.32	0.000	4	532
KLK5	kallikrein related peptidase 5	37	y	−2.57	0.000	4	2,007
KLK14	kallikrein related peptidase 14	64	y	−1.51	0.012	1	6,276
KLK4	kallikrein related peptidase 4	76		0.70	0.158		1,682
KLK6	kallikrein related peptidase 6	105	y	−2.30	0.000	4	1,787
KLK15	kallikrein related peptidase 15	125					3,921
KLK10	kallikrein related peptidase 10	136	y	−2.14	0.001	5	6,069
KLK11	kallikrein related peptidase 11	152		−0.37	0.602		3,035
KLK1	kallikrein 1	163		−0.50	0.260		1,956
KLK2	kallikrein related peptidase 2	270					2,082
KLKB1	kallikrein B1	391		0.56	0.082		1,846
KLK3	kallikrein related peptidase 3	1,012					3,992
KLK9	kallikrein related peptidase 9	10,373					18,931

Table B.4: Summary of Ranked Kallikreins for BTNR using GeneSet MAPR, Part 2 of 2

Gene Name	MAPR rank	BTNR in set	SNV nR	SNV pCR	CNV nR	CNV pCR	TCGA, BC	TCGA, all	COSMIC Tier
KLK13	20	y						0.011	
KLK8	28	y						0.014	
KLK12	30							0.013	
KLK7	35	y						0.008	
KLK5	37	y						0.013	
KLK14	64	y	0.000	0.023				0.007	
KLK4	76							0.014	
KLK6	105	y						0.013	
KLK15	125							0.016	
KLK10	136	y						0.008	
KLK11	152							0.012	
KLK1	163		0.000	0.023					
KLK2	270							0.012	1
KLKB1	391							0.016	
KLK3	1,012								
KLK9	10,373								

Table B.5: Summary of Ranked Collagen Type N Alpha Chains for BTNR using GeneSet MAPR, Part 1 of 2

Gene Name	Description	MAPR rank	BTNR in set	BTNR logFC	BTNR p-val	Others in set	Others med rnk
COL20A1	col. type XX α1 chain	25					3,014
COL9A3	col. type IX α3 chain	40	y	1.38	0.006	4	329
COL9A2	col. type IX α2 chain	48		0.71	0.008	1	497
COL27A1	col. type XXVII α1 chain	55		0.95	0.000	1	724
COL28A1	col. type XXVIII α1 chain	57	y	0.57	0.234		449
COL22A1	col. type XXII α1 chain	59		−1.05	0.054	1	1,317
COL6A6	col. type VI α6 chain	62		−0.32	0.462		1,823
COL6A5	col. type VI α5 chain	67	y	−2.02	0.003	1	2,183
COL24A1	col. type XXIV α1 chain	72		0.54	0.120		676
COL10A1	col. type X α1 chain	88		0.65	0.186	1	376
COL12A1	col. type XII α1 chain	89		−0.21	0.585		258
COL25A1	col. type XXV α1 chain	95		−0.06	0.864	1	1,380
COL4A1	col. type IV α1 chain	96		−0.07	0.776		658
COL4A4	col. type IV α4 chain	98		−0.38	0.200		1,586
COL4A5	col. type IV α5 chain	100		0.01	0.966	3	824
COL21A1	col. type XXI α1 chain	106		0.23	0.543		1,776
COL9A1	col. type IX α1 chain	113	y	1.86	0.002	2	735
COL4A6	col. type IV α6 chain	120	y	0.75	0.088	1	1,395
COL23A1	col. type XXIII α1 chain	138		−0.44	0.182		2,543
COL14A1	col. type XIV α1 chain	155		0.38	0.351	2	864
COL8A2	col. type VIII α2 chain	156		−0.08	0.781		1,268
COL6A3	col. type VI α3 chain	166		0.04	0.885	3	2,386
COL11A1	col. type XI α1 chain	170		0.09	0.871	2	208
COL6A2	col. type VI α2 chain	171		−0.41	0.069	1	1,016
COL17A1	col. type XVII α1 chain	185		0.09	0.870		197
COL5A3	col. type V α3 chain	191		0.13	0.644		711
COL5A2	col. type V α2 chain	198		−0.03	0.918	2	2,157
COL2A1	col. type II α1 chain	233	y	1.74	0.020	2	646
COL1A2	col. type I α2 chain	267		−0.05	0.888	2	730
COL13A1	col. type XIII α1 chain	271		0.14	0.629		2,581
COL16A1	col. type XVI α1 chain	274		0.13	0.665	1	1,615
COL6A1	col. type VI α1 chain	278		−0.11	0.633		1,249
COL18A1	col. type XVIII α1 chain	282		−0.03	0.881	1	459
COL4A2	col. type IV α2 chain	340		−0.11	0.592	2	1,458
COL5A1	col. type V α1 chain	348		−0.22	0.499	1	658
COL3A1	col. type III α1 chain	378		0.05	0.869	2	525
COL4A3	col. type IV α3 chain	390		−0.06	0.849		1,383
COL15A1	col. type XV α1 chain	402		0.35	0.233		755
COL1A1	col. type I α1 chain	411		−0.06	0.854	3	627
COL7A1	col. type VII α1 chain	419		−0.12	0.732	1	693
COL8A1	col. type VIII α1 chain	472		0.56	0.210		1,701
COL26A1	col. type XXVI α1 chain	734					3,512
COL19A1	col. type XIX α1 chain	1,103		−0.51	0.378		7,225
COL11A2	col. type XI α2 chain	3,108					7,185

(Omitted: 3 items of higher rank)

Table B.6: Summary of Ranked Collagen Type N Alpha Chains for BTNR using GeneSet MAPR, Part 2 of 2

Gene Name	MAPR rank	BTNR in set	SNV nR	SNV pCR	CNV nR	CNV pCR	TCGA, BC	TCGA, all	COSMIC Tier
COL20A1	25							0.029	
COL9A3	40	y						0.018	
COL9A2	48							0.018	
COL27A1	55							0.035	
COL28A1	57	y						0.027	
COL22A1	59		0.011	0.000	0.068	0.364	0.028	0.071	
COL6A6	62						0.026	0.059	
COL6A5	67	y					0.036	0.040	
COL24A1	72							0.039	
COL10A1	88							0.013	
COL12A1	89		0.023	0.023			0.034	0.065	
COL25A1	95							0.026	
COL4A1	96							0.042	
COL4A4	98						0.021	0.048	
COL4A5	100		0.011	0.000			0.028	0.048	
COL21A1	106		0.000	0.023				0.032	
COL9A1	113	y						0.034	
COL4A6	120	y					0.019	0.037	
COL23A1	138							0.014	
COL14A1	155		0.000	0.045	0.034	0.182	0.030	0.050	
COL8A2	156							0.011	
COL6A3	166		0.011	0.000			0.034	0.083	
COL11A1	170		0.023	0.000			0.019	0.081	
COL6A2	171							0.036	
COL17A1	185		0.000	0.023				0.032	
COL5A3	191		0.023	0.000			0.019	0.043	
COL5A2	198		0.000	0.023			0.017	0.043	
COL2A1	233	y	0.011	0.000				0.034	1
COL1A2	267		0.000	0.045			0.018	0.047	
COL13A1	271								
COL16A1	274		0.000	0.023				0.032	
COL6A1	278		0.000	0.023				0.024	
COL18A1	282						0.019	0.030	
COL4A2	340							0.038	
COL5A1	348		0.000	0.023			0.023	0.059	
COL3A1	378		0.011	0.023				0.047	2
COL4A3	390								
COL15A1	402								
COL1A1	411							0.031	1
COL7A1	419		0.000	0.023			0.023	0.055	
COL8A1	472								
COL26A1	734							0.011	
COL19A1	1,103						0.018	0.042	
COL11A2	3,108								

(Omitted: 3 items of higher rank)

APPENDIX C

COMPARISON OF TERM ENRICHMENT FOR BTNR BEFORE AND AFTER GENESET MAPR

Table C.1 contains annotation terms which were in the top 100 terms returned by GSEA, after running GeneSet MAPR on the BTNR set. Terms were filtered out that had a median rank of 100 or better over the other 9 compared gene sets, as these were considered broadly applicable to cancer sets and therefore less interesting. The first three columns indicate the annotation term name, brief description, and size. The next three show the rank returned from GSEA when using the original BTNR set, the MAPR similarity ranking, and the median rank over 9 other cancer sets. The last four columns show, for each annotation term, the proportion of genes from the group indicated in the row header that appear in that term. The four groups are the Claudin family, the Kallikrein family, the Collagen Type N Alpha Chain family, and the top 323 genes as ranked by MAPR.

Table C.1: Top 100 Enriched Annotation Terms from GSEA after Using GeneSet MAPR, Omitting Those Common to Other Gene Sets

Term	Description	Size	BTNR rank	MAPR rank	Med rank	CLD	KLK	COL	MAPR 323
GO:0004252	serine-type endopeptidase activity	241	1,112	3	215	–	0.93	–	0.17
PF00089.21	Trypsin	123	768	4	265	–	1.00	–	0.16
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	160	421	5	236	–	–	–	0.15
PF00028.12	Cadherin	116	365	7	392	–	–	–	0.14
PF00008.22	EGF	171	505	8	106	–	–	–	0.08
PF08266.7	Cadherin_2	96	661	9	464	–	–	–	0.13
PF09342.6	DUF1986	94	744	10	321	–	0.93	–	0.13
GO:0034765	regulation of ion transmembrane transport	118	1,720	11	441	–	–	–	0.03
GO:0006811	ion transport	167	532	13	286	0.13	–	–	0.02

Continued on next page →

Table C.1 – continued from previous page

Term	Description	Size	BTNR rank	MAPR rank	Med rank	CLD	KLK	COL	MAPR 323
PF00520.26	Ion_trans	121	856	14	514	–	–	–	0.02
PF13365.1	Trypsin_2	80	645	15	414	–	0.73	–	0.11
PF01391.13	Collagen	87	1,096	17	291	–	–	0.94	0.08
GO:0071805	potassium ion transmembrane transport	136	1,947	18	597	–	–	–	0.00
PF07645.10	EGF_CA	100	718	19	188	–	–	–	0.06
PF07885.11	Ion_trans_2	81	1,505	20	549	–	–	–	0.01
GO:0030574	collagen catabolic process	75	684	21	162	–	0.07	0.81	0.09
PF12947.2	EGF_3	88	1,017	22	245	–	–	–	0.05
GO:0004222	metalloendopeptidase activity	135	3,197	23	475	–	–	–	0.06
GO:0007268	chemical synaptic transmission	420	190	24	134	–	–	–	0.02
PF12662.2	cEGF	102	469	26	179	–	–	–	0.04
PF13583.1	Reprolysin_4	59	1,198	28	384	–	–	–	0.03
path:map04974	Protein digestion and absorption	90	1,635	29	278	–	–	0.74	0.07
GO:0008201	heparin binding	157	442	30	116	–	–	0.09	0.03
PF07690.11	MFS_1	136	1,899	31	1,006	–	–	–	0.01
PF13582.1	Reprolysin_3	58	1,488	32	468	–	–	–	0.03
PF12661.2	hEGF	107	1,221	33	342	–	–	–	0.03
PF13519.1	VWA_2	69	1,730	34	808	–	–	0.26	0.05
PF00041.16	fn3	193	251	35	312	–	–	0.09	0.02
PF02210.19	Laminin_G_2	62	1,700	36	400	–	–	0.28	0.03
PF00092.23	VWA	71	1,252	37	668	–	–	0.26	0.05
GO:0004867	serine-type endopeptidase inhibitor activity	102	–	38	301	–	–	0.06	0.06
path:map04080	Neuroactive ligand-receptor interaction	276	1,133	39	142	–	–	–	0.01
PF13574.1	Reprolysin_2	52	1,382	40	476	–	–	–	0.02
GO:0005604	basement membrane	79	1,059	41	204	–	–	0.19	0.05
GO:0070588	calcium ion transmembrane transport	153	478	42	489	–	–	–	0.01
GO:0035725	sodium ion transmembrane transport	106	1,946	43	1,247	–	–	–	0.02
GO:0008076	voltage-gated potassium channel complex	93	1,675	44	579	–	–	–	–
path:map04514	Cell adhesion molecules (CAMs)	141	710	45	203	0.92	–	–	0.04
PF13688.1	Peptidase_M84	54	1,614	46	445	–	–	–	0.02
GO:1902476	chloride transmembrane transport	116	1,601	47	808	–	–	–	0.01
PF07974.8	EGF_2	105	929	48	294	–	–	–	0.03
PF08016.7	PKD_channel	59	1,774	49	1,313	–	–	–	0.02
PF00413.19	Peptidase_M10	46	1,581	50	356	–	–	–	0.03
GO:0010951	negative regulation of endopeptidase activity	202	–	51	311	–	–	0.09	0.07

Continued on next page →

Table C.1 – continued from previous page

Term	Description	Size	BTNR rank	MAPR rank	Med rank	CLD	KLK	COL	MAPR 323
GO:0006810	transport	408	224	52	297	–	–	–	0.03
GO:0005581	collagen trimer	70	1,095	53	588	–	–	0.53	0.06
GO:0005201	extracellular matrix structural constituent	60	1,773	54	301	–	–	0.45	0.07
GO:0005796	Golgi lumen	88	1,964	55	410	–	–	–	0.01
unmapped	C2_set_2	226	2,832	56	281	–	–	–	0.01
PF13768.1	VWA_3	44	2,616	57	939	–	–	0.23	0.05
GO:0008236	serine-type peptidase activity	59	2,511	58	1,047	–	0.53	–	0.04
PF00822.15	PMP22_Claudin	60	1,333	59	2,631	1.00	–	–	0.06
PF00083.19	Sugar_tr	74	2,831	60	1,699	–	–	–	0.00
PF02480.11	Herpes_gE	151	739	61	666	–	–	–	0.04
PF00090.14	TSP_1	62	1,098	62	508	–	–	–	0.02
PF13903.1	Claudin_2	82	1,426	64	3,371	1.00	–	–	0.06
PF13306.1	LRR_5	107	1,091	66	921	–	–	–	0.01
PF13385.1	Laminin_G_3	65	1,632	67	527	–	–	0.28	0.03
PF07648.10	Kazal_2	51	2,897	68	1,876	–	–	–	0.03
GO:0005178	integrin binding	104	406	69	148	–	–	0.09	0.02
GO:0001501	skeletal system development	136	321	70	228	–	–	0.21	0.03
PF00046.24	Homeobox	246	2,715	71	203	–	–	–	–
PF01400.19	Astacin	26	2,748	72	384	–	–	–	0.03
PF01462.13	LRRNT	73	2,507	73	1,061	–	–	–	0.01
GO:0045211	postsynaptic membrane	210	1,025	74	417	–	–	–	0.02
GO:0006814	sodium ion transport	88	1,299	75	1,618	–	–	–	0.01
GO:0034220	ion transmembrane transport	293	273	76	248	–	–	–	0.01
PF01421.14	Reprolysin	41	2,533	77	1,081	–	–	–	0.01
PF01562.14	Pep_M12B_propep	39	2,557	78	1,089	–	–	–	0.01
PF00050.16	Kazal_1	49	3,615	79	2,157	–	–	–	0.03
GO:0016323	basolateral plasma membrane	186	409	80	273	0.13	–	–	0.02
GO:0006813	potassium ion transport	88	2,206	81	685	–	–	–	0.01
GO:0016324	apical plasma membrane	274	377	82	156	0.08	–	–	0.02
GO:0005245	voltage-gated calcium channel activity	50	3,216	83	1,337	–	–	–	0.01
PF00054.18	Laminin_G_1	44	2,822	85	503	–	–	0.09	0.01
PF13520.1	AA_permease_2	28	918	86	3,441	–	–	–	0.01
GO:0005254	chloride channel activity	64	3,007	87	1,563	–	–	–	0.01
GO:0005249	voltage-gated potassium channel activity	63	4,199	88	1,061	–	–	–	0.00
GO:0030054	cell junction	444	142	89	268	–	–	–	0.02
PF02932.11	Neur_chan_memb	57	2,249	90	1,240	–	–	–	0.01
GO:0005518	collagen binding	59	1,788	91	429	–	–	0.04	0.01

Continued on next page →

Table C.1 – continued from previous page

Term	Description	Size	BTNR rank	MAPR rank	Med rank	CLD	KLK	COL	MAPR 323
PF01049.12	Cadherin_C	29	2,443	92	839	–	–	–	0.01
GO:0016338	calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules	21	462	94	1,883	0.88	–	–	0.05
GO:0042391	regulation of membrane potential	78	1,500	96	936	–	–	–	0.02
PF01094.23	ANF_receptor	39	2,654	97	1,303	–	–	–	0.00
path:map04610	Complement and coagulation cascades	69	1,857	99	165	–	0.07	–	0.02
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	47	3,042	100	583	–	–	–	0.01

End of table.